University of Cape Town

The Evaluation of Case-Mix Adjusted Efficiency Scores: The Case of the South African Private Hospital Industry

Kathryn Ann Dreyer
DRYKAT001

Dissertation submitted in full fulfilment of the requirements for the degree of Master of Business Science in Actuarial Science

Faculty of Commerce

May 22, 2013

Abstract

There is little existing South African literature relating to hospital efficiency that allows for differences in case mix across hospitals. One of the primary motivations for this dissertation is to help fill this gap in the literature by examining the impact that adjusting for differences in case mix has on efficiency scores.

Data Envelopment Analysis (DEA) is chosen as the efficiency measurement method because of its flexibility and ease of handling multiple inputs and outputs. A number of DEA models are applied to a sample of South African private hospitals for the years 2008 to 2011 inclusive.

Three different case-mix adjustment techniques are investigated and their ability to capture differences in case mix is assessed. The three techniques investigated are: a case-mix adjustment factor (constructed using Diagnosis-Related Groups (DRGs)) to adjust outputs; including the case-mix adjustment factor as an additional output; and disaggregating hospital admissions into broad categories which are used as outputs.

A comparison of the unadjusted model with the case-mix adjusted model reveals that omitting the adjustment can have a considerable impact on efficiency scores. Whilst little difference is noted in average efficiency scores for the group of hospitals, 90% for the unadjusted model and 92% for the adjusted model in 2011, there are substantial differences between the adjusted and unadjusted efficiency scores of individual hospitals.

On comparison of the three different techniques investigated, it is evident that if there is sufficient data to construct a case-mix adjustment factor, case-mix adjusted admissions should be used, rather than using the factor as an additional output variable. In the case where insufficient data is available, disaggregating admissions does capture some of the differences in case mix but a substantial amount of power is lost as a result of increasing the number of output variables.

It is noted that a more efficient hospital industry is necessary in order to progress with health care reform in South Africa and this dissertation is a useful first step to evaluating, managing and improving hospital efficiency in South Africa.

Declaration

I hereby declare that:

1. this is my own unaided work, and that each significant contribution to, and quotation in, this dissertation from the work of other people has been cited and referenced.

2. neither the substance nor any part of the thesis has been submitted in the past, or is being, or is to be submitted for a degree at this University or any other University.

I grant the University of Cape Town free license to reproduce for the purpose of research either the whole or any portion of the contents.

K Dreyer

February 2013

Acknowledgements

The inspiration and determination to do my masters came from my friend and supervisor, Shivani Ramjee. Her constant itch to learn more, unfaltering support and her warm words of encouragement helped me in a myriad of ways.

A special thanks to Andrew Linden, friend and fellow master's student, for the hours of extracting, cleaning and analysing data together and for his sense of humour that encouraged me to just keep going. I would also like to thank my parents, Craig and Linda Dreyer, whose support and encouragement gave me the impetus to finish writing.

To all the people who read and reread the numerous drafts, Mary Derman, Szymon Marszalek and Frans Cronje, your feedback and moral support was invaluable.

My sincere thanks to everyone at the hospital group who collected and provided the data and who were so willing to answer all my questions.

The financial assistance of the National Research Foundation (DAAD-NRF) and the Institute of Applied Statistics towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the DAAD-NRF or the Institute of Applied Statistics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

The delivery of hospital services, and health care services in general, in South Africa is characterised by a two-tier system, namely the public sector and the private sector. The public sector is government funded and is designed to offer universal coverage in that public health care is available to the whole population on a means-tested basis (Ramjee and McLeod, 2010). The private sector is funded predominantly by medical schemes[1] and private out-of-pocket expenditure (Ramjee and McLeod, 2010). This two-tier system is perceived as being extremely inequitable across the two sectors, public and private, in terms of both financing and delivery of health care (Wadee et al., 2003). Furthermore, large inefficiencies are perceived to be present within each of the two hospital sectors (Wadee et al., 2003). This is supported by high expenditure on health care and comparatively poor health outcomes (WHO, 2009; Department of Health, 2011b). The South African health care system is currently under review in order to "promote equity and efficiency so as to ensure that all South Africans have access to affordable, quality health care services regardless of their socio-economic status" (Department of Health, 2011a). The need to quantify and improve efficiency within a hospital context will become increasingly important in order to meet these objectives.

This piece of research considers the issues of measuring differences in the types of cases treated across hospitals, known as case mix, and the impact that adjusting for case

[1]Medical Schemes are not-for-profit entities, contributions to which are tax-subsidised. The business of a medical scheme is defined in the Department of Health (1998) as "the business of undertaking liability in return for a premium or contribution (a) to make provision for the obtaining of any relevant health service; (b) to grant assistance in defraying expenditure incurred in connection with the rendering of any relevant health service; and (c) where applicable, to render a relevant health service, either by the medical scheme itself, or by any supplier or group of suppliers of a relevant health service or by any person, in association with or in terms of an agreement with a medical scheme".

mix has on the measurement of technical efficiency. A number of methods for measuring case mix in the context of measuring hospital efficiency are investigated. Application of these methods is demonstrated through measuring the efficiency of a group of South African private hospitals using historical operational, clinical and human resource data to construct a Data Envelopment Analysis (DEA) model.

This dissertation examines the efficiency of the private hospital sector. There are three major reasons for this choice. Firstly, with the proposed introduction of the National Health Insurance (NHI), private hospitals may wish to contract with the NHI in order to expand their patient base and, in this event, are likely to need to be able to demonstrate that they are efficient for pricing purposes. Although DEA efficiency scores are relative measures of efficiency, this is a valuable tool to identify poorly performing hospitals and which will help improve the overall group efficiency. Secondly, the private hospital industry is under pressure to manage increasing hospital costs and efficiency measurement and management may be able to help ensure that hospitals are sustainable in the face of pricing pressure. Thirdly, whilst it would have been ideal to use data from both the public and private sector, the lack of comprehensive information systems in public hospitals means that the data required are either incomplete or unavailable for many public hospitals. One of the aims of this dissertation is to develop the skills and capacity required to measure hospital efficiency, such that the methodology can, in the future, be applied to any hospital environment and to provide additional insight into 'best practice' in hospital efficiency analysis. The lack of detailed data of sufficient quality and the need to develop these skills was highlighted in two major studies conducted in South African by Kibambe and Koch (2007) and Zere, McIntyre and Addison (2001).

The development of an efficiency methodology and the results of measuring hospital efficiency in South Africa could be used by policymakers, hospital management teams, as well as shareholders. Efficiency scores could be used by policymakers during the process of price-setting and negotiating between providers and financers of health care. The ability to identify efficient hospitals can be used by hospital management in order to identify successful management strategies and transfer these strategies to those hospitals that are not fully efficient. Shareholders will be interested in efficiency scores from an investment point of view.

1.2 Objectives

The overall aim of this dissertation is to investigate the importance of adjusting for differences in case mix across hospitals, when measuring the relative efficiency of hospitals.

More specifically, the major aims of this dissertation are:

1. To provide a definition of efficiency and an introduction to Data Envelopment Anal-

ysis (DEA), as an efficiency measurement technique.

2. To explore the different methods available to adjust for differences in case mix across hospitals.

3. To determine whether adjusting for case mix has an impact on efficiency scores.

4. To examine the different techniques and the impact each has on efficiency scores by applying DEA to a set of South African private hospitals and to determine which of the techniques is most appropriate.

1.3 Scope and Limitations

The focus of this dissertation is the South African private hospital industry. Therefore, the results drawn from this dissertation cannot be applied to and may not be representative of public hospitals in South Africa.

It is also important to note that the sample used is drawn from only one of the three major hospital groups in South Africa. So, although the sample is large and represents approximately 25% of the private hospital industry, the results and conclusions may not be representative of the private hospital industry as a whole, rather they relate specifically to one hospital group. Whilst the results are not applicable to the other hospital groups, the same methodology could be applied. The three main hospital groups are structurally very similar to one another, in terms of the inputs required to produce the outputs. This implies that the methodology used in this piece of research, as opposed to the results themselves, could be applied to the other hospital groups.

Technical efficiency is the only component of efficiency investigated in this dissertation. The motivation for this is that technical efficiency analyses the amount of output produced, relative to the input used. The relationship between inputs and outputs is expected to change when an adjustment for case mix has been made.

The results of this paper are *relative* efficiency scores and are not an indication of absolute efficiency. This is one of the limitations of DEA models.

1.4 Overview

This dissertation consists of eight chapters. The current chapter provides an introduction to the topic and contextualises the investigation. In the following two chapters, an overview of the private hospital industry is presented, along with its potential role in relation to National Health Insurance (NHI).

This is followed by an exploration of efficiency, its definition and the different components that exist. Before a method for measuring efficiency is selected, the different techniques available are investigated and the reasoning behind selecting Data Envelopment Analysis (DEA) as the preferred method is explored. DEA is discussed in detail, as well as its mathematical construction, before moving on to the applications to the South African hospital industry.

Adjusting for case mix is central to this dissertation, therefore a clear understanding of case mix is necessary. The various techniques available to adjust for case mix are investigated.

Since data are key in any investigation, the next chapter covers the source, characteristics and quality of the data used and explains the reasoning behind the the major methodological decisions.

Finally, results of the different models are analysed before arriving at conclusions for this research.

Chapter 2

The Private Hospital Industry

2.1 Introduction

The aim of this chapter is to provide an overview of the private hospital sector in South Africa and to discuss the potential roles that the private hospital industry can play in a National Health Insurance (NHI) environment.

A comprehensive understanding of the private hospital environment is essential for this piece of research, as it contextualises the research question, as well as demonstrates the importance of measuring the efficiency of hospitals in an uncertain and changing environment. Furthermore, a good understanding of the hospital environment is necessary in order to inform decisions regarding the inputs and outputs of the model, as well as other structural decisions.

2.2 The Existing Private Hospital Industry

The private hospital industry in South Africa is examined from a number of different perspectives including the growth of the industry, ownership, financing, the number of beds (which is used as a proxy for the size of a hospital) and human resource constraints of the industry. There are limited sources on this subject and majority of the statistics below have been taken from Matsebula and Willie (2007). Where possible, this has been supplemented with other references.

The private hospital industry experienced periods of strong growth, in particular during the early 1990's (Söderland, Schierhout and van der Heever, 1998). This corresponds to a period where there was a distinct shift by medical scheme beneficiaries away from public hospitals and an increase in demand for private care (Söderland et al., 1998). The growth in the private hospital industry continued from the late 1990's. Private hospital beds increased by approximately 32% between 1998 and 2006 (Söderland et al., 1998;

Matsebula and Willie, 2007). This growth coincides with a period during which the number of hospital beds in the public sector was reduced in almost every province (Boulle, Blecher and Burn, 2008). The closure of public hospitals, a change in public perceptions of the quality of public sector hospital care and the preference for private hospital treatment by the medical scheme beneficiary population all had a positive impact on the growth of the private hospital industry (Matsebula and Willie, 2007).

Currently, the private hospital sector is dominated by three hospital groups: Life Healthcare, Mediclinic and Netcare. Collectively, these three hospital groups own 66.5% of all the private hospitals in South Africa (Matsebula and Willie, 2007). The majority of the remainder of the private hospitals are affiliated to the National Hospital Network (NHN). The objective of the NHN is to bring together these independently owned hospitals in order to create synergistic relationships and to assist these hospitals to achieve efficient and effective delivery of health care services to their patients (National Hospital Network, 2012). Private hospitals are for-profit institutions and have been made consistent profits over the years (Matsebula and Willie, 2007). All three hospital groups are registered on the Johannesburg Stock Exchange (JSE) and they experienced consistent growth in earnings prior to the recession in 2008 (Matsebula and Willie, 2007).

The concentration of hospital ownership has been blamed for escalating medical expenses and hospital costs in the private sector. All three hospital groups have been involved in large mergers that have been taken to the Competition Tribunal (Competition Tribunal, 2010, 2006, 2005). It has been suggested that through these large mergers the hospital groups have used their power and market share to influence the price of hospital services (Matsebula and Willie, 2007). However, the private hospital industry has disputed these claims. The substantial increase in costs in the early 2000's has been blamed on the near 60% collapse of the ZAR, which significantly increased the cost of supplies (Matsebula and Willie, 2007). Furthermore, the annual escalation of inpatient days coupled with a deteriorating risk profile of the insured population due to changes in the policy environment, as a result of the Medical Schemes Act, No 131 of 1998, has also contributed to increasing costs of hospital care (Matsebula and Willie, 2007).

Medical Schemes are the main source of revenue for private hospitals which provide care predominantly for the medical scheme beneficiaries. As a result, the viability of private hospitals is dependent *inter alia* on the medical scheme environment (Matsebula and Willie, 2007). However, the occurrence of out-of-pocket expenditure on private hospitals has been increasingly occurring in the uncovered population (Matsebula and Willie, 2007).

The three major groups collectively own and operate 75.6% of the private sector beds and 80.0% of theatre facilities in the private sector (Matsebula and Willie, 2007). Private hospital beds are distributed in metropolitan areas with the distribution of private hospitals mirroring the distribution of medical scheme beneficiaries across South Africa (Matsebula and Willie, 2007; Council for Medical Schemes, 2012). All three major private hospital groups offer very similar services. The difference among the three groups in respect of access to surgical theatres, 24 hour emergency care, catheterisation laboratories and MRI scanners is negligible (Matsebula and Willie, 2007). That is, the three hospital groups are structurally almost identical to one another. This is particularly important to note, as it means that the methodology presented in this dissertation can be applied to any one of these hospital groups.

Due to the for-profit nature of private hospitals, the industry has been able to develop and establish a range of facilities that would otherwise not be available due to competing priorities in the public sector (Matsebula and Willie, 2007).

As per the ethical rules of the Health Professions Council of South Africa, private hospitals are not allowed to employ doctors and other health professionals, with the exception of nurses (Health Professions Council of South Africa, 1974). The relationship between doctors and hospitals is a mutual one. Doctors are indirect sellers of the hospital services, whilst the doctors are dependent on hospitals to provide a complete service to the patients (Matsebula and Willie, 2007). As a result of doctors billing patients for care received, rather than being employed by the hospital, there is a risk of supplier induced demand and agency problems. In other words, there is a risk that doctors suggest unnecessary care and medication as they are often more knowledgeable than patients and they benefit financially from selling their services.

Since doctors cannot be directly employed, private hospitals invest heavily in infrastructure to attract doctors to particular hospitals (Matsebula and Willie, 2007). This method of attracting and retaining medical practitioners has been criticised for pushing up costs of hospital care, as well as impacting negatively on the public sector by drawing health professionals out of it and into the private sector, exacerbating inequity between the two sectors (Matsebula and Willie, 2007). This strategy is also expensive, as the cost of medical equipment is substantial and can often lead to duplication. For example, magnetic resonance image (MRI) scanners, of which there are a large number in the private sector (Matsebula and Willie, 2007), are often underutilised and may result in allocative inefficiencies in the industry. It has been suggested that as a result of non-price competition and the strategy of attracting good staff, hospitals have been over-capitalised (Council for Medical Schemes, 2008).

Econex (2010) highlight the shortage of nurses in South Africa, which is affecting both the public and the private sectors. This is problematic as hospitals rely heavily on nurses to successfully run hospitals (Matsebula and Willie, 2007). Private hospitals have attributed escalating medical costs partially to the limited supply of nurses, as hospitals are forced to compete for nursing staff through more attractive remuneration packages

(Matsebula and Willie, 2007).

All of the above factors offer insight into the existing private hospital industry.

2.3 Private Hospitals in a National Health Insurance Environment

The existing two-tier South African health care system is currently being reviewed in order to "promote equity and efficiency so as to ensure that all South Africans have access to affordable, quality health care services regardless of their socio-economic status" (Department of Health, 2011a). The concept underlying National Health Insurance (NHI) is that it should be a system that is universal, compulsory and free at the point of use (Centre for Development and Enterprise, 2011), as well as being equitable and sustainable in the long term (Department of Health, 2011a).

At the launch of the Green Paper on NHI, Aaron Motsoaledi, Minister of Health, summarised that one of the challenges and aims of the NHI is to draw on the strengths of both the public and the private health care sectors (Centre for Development and Enterprise, 2011). One of the proposed interactions between the NHI fund and private hospitals is that the NHI fund will purchase health care services from private hospitals.

The private health care sector delivers health care services competently, but at prices which result in only a very small proportion of the population being able to afford these services (Centre for Development and Enterprise, 2011). In order to effectively contract with the NHI fund, private hospitals may need to review their levels of efficiency.

The Centre for Development and Enterprise (2011) makes three other recommendations regarding the ways in which the private sector is able to support health care reform:

1. Create opportunities for private sector medical practitioners and specialists to work in the public sector.

2. Encourage and extend public-private partnerships from infrastructure only, to include hospital management, supply chain management and clinical services.

3. Develop a joint public/private plan on health professionals' needs such that the training of these health professionals can be facilitated by the private sector by easing regulations.

These recommendations are currently being investigated for feasibility.

A shift from the existing two-tier system to an NHI environment is likely to have a significant impact on the current profile of patients treated, as well as the pattern of resource consumption, in the private sector and the public sector respectively. It is important to understand the existing profile of patients within each sector in order to fully understand how these profiles will change when health care reform occurs. Measuring case mix is an important aspect of this.

Chapter 3

Defining and Measuring Efficiency

3.1 An Introduction to Efficiency

Efficiency is a well established concept in the field of economics (Nguyen and Coelli, 2009). Nguyen and Coelli (2009) provide a broad definition of efficiency and define it as how well an organisation is able to achieve its goals by converting resources, or inputs, into outputs, given the technology available. Resources consist of all the factors of production and include, *inter alia*, capital investment, raw materials and labour (Nguyen and Coelli, 2009).

Confusion arises over the definition of efficiency as a result of the different components that exist. For example, efficiency could be thought of as an organisation using the minimum level of inputs to produce a given set of outputs. Alternatively, it could be thought of as the ability of the organisation to produce outputs at the lowest cost. Whilst neither of these definitions is strictly incorrect, they do not offer a holistic view of the overall efficiency of an organisation. Rather, these definitions only deal with particular components of the concept of efficiency.

Farrell (1957) proposed that the overall efficiency of an organisation consisted of two major components; namely, technical efficiency and allocative efficiency. Overall efficiency is also known as *cost efficiency*, *economic efficiency* and *productive efficiency.* Sherman and Zhu (2006) and Coelli, Rao, O'Donnell and Battese (2005) expand on Farrell's definition by adding scale efficiency as one of the components. Whilst other components of efficiency exist, for example financing efficiency which relates to the efficient use of equity, these four components encompass other possible components and are widely accepted in economic literature (Sherman and Zhu, 2006). Efficiency analysis is a multi-faceted concept and each component of efficiency requires careful management in order for an organisation to be fully efficient.

Efficiency is particularly difficult to define within a health care context, specifically within the hospital environment, because of the complexity of the organisations (Sherman and Zhu, 2006) and the difficulty in defining what constitutes an input and an output (Cook and Zhu, 2007).

3.2 Components of Efficiency

3.2.1 Technical Efficiency

Technical efficiency, also known as productive efficiency, was first analysed by Debreu (1951). Debreu's measurement aimed to determine the wastage associated with a non-optimal situation. Farrell (1957) expanded on Debreu's definition and developed a measure for technical inefficiency which formed the basis of existing efficiency analysis. Farrell (1957) defines an organisation as technically efficient when the greatest number of outputs are produced from a given set of inputs. Alternatively, an organisation is technically efficient when using the minimum level of inputs required to produce a given set of outputs.

Coelli et al. (2005) and Hollingsworth, Dawson and Maniadakis (1999) define a technically efficient organisation as one that is operating on the production frontier. The production frontier defines the relationship between the inputs and outputs and represents the maximum output that can be produced from a given set of inputs.

In the hospital environment, technical efficiency refers to the relationship between resource utilisation (for example pharmacy goods, theatre time and doctors visits) and health outcomes (such as the number of patients treated, total days spent in hospital, decrease in mortality rates or increased life expectancy) (Worthington, 2004). Nguyen and Coelli (2009) provide a similar definition of efficiency in the hospital environment.

3.2.2 Allocative Efficiency

Allocative efficiency refers to an organisation utilising the optimal combination of inputs, given current input prices and the current available technology, to produce the maximum possible outputs (Coelli et al., 2005; Worthington, 2004; Linna, 1998). This optimal mix of inputs, with respect to price, will be constrained by the required quality standard of inputs (Sherman and Zhu, 2006). Allocative efficiency can similarly be defined as producing an optimal mix of outputs, given the prices of outputs and the available technology.

In order to estimate allocative efficiency, price information is required, as well as an assumption regarding cost minimising or profit maximising behaviour of the organisation (Nguyen and Coelli, 2009). Deciding upon this behavioural assumption becomes difficult in the context of the health care environment. Neither assumption may be appropriate for public hospitals that are focused on providing quality care for as many patients as possible. However, when there are significant budget constraints, the assumption of cost minimisation may be appropriate. In the South African private hospital sector, in contrast, the profit maximisation assumption is appropriate as private hospitals are for-profit organisations.

3.2.3 Cost Efficiency

An organisation which is cost efficient is defined to be both technically and allocatively efficient (Coelli et al., 2005). Thus, cost efficiency is a measure of overall efficiency. Nguyen and Coelli (2009) define cost efficiency to be the extent to which the objectives of an organisation are met, relative to the economic resources used. In order to improve the cost efficiency, an organisation would aim to improve either allocative or technical efficiency through using a different combination of inputs or by using fewer inputs. Alternatively, for a given set of inputs, an organisation could aim to increase the number of outputs or change the combination of outputs produced to improve efficiency. The exact relationship between technical, allocative and cost efficiency is expanded upon in Section 3.3.1.

3.2.4 Scale efficiency

Banker, Charnes and Cooper (1984) describe an organisation as being scale efficient if it is operating at the optimal size given the relative mix of inputs and outputs. Coelli et al. (2005) define an organisation as scale efficient if it is operating on the point of the production frontier that maximises productivity. Organisations which are producing more or fewer goods than the optimal level, given the relative mix of inputs and outputs, will experience increased marginal costs (Sherman and Zhu, 2006). This should not be confused with economies of scale which looks exclusively at the cost advantages arising due to a change in size (Sloman, 2006).

There are three marked types of returns to scale (Coelli et al., 2005). Decreasing returns to scale occur when a proportionate increase in inputs results in a less than proportionate increase in outputs. Furthermore, when an organisation produces more goods than the optimal level, it is said to be operating under decreasing returns to scale. Similarly, an organisation which produces fewer outputs than the optimal level is said to be operating under increasing returns to scale (Coelli et al., 2005). Under increasing returns to scale, a proportionate increase in inputs will result in a more than proportionate increase in outputs. An organisation operating under constant returns to scale will experience an equally proportionate increase in outputs when inputs are increased proportionately.

A production frontier can exhibit different return to scale properties. When it displays more than one type of returns to scale, it is known as a variable returns to scale (VRS) production frontier. This is illustrated in Figure 3.2, in Section 3.3.2.

In a hospital environment, scale inefficiencies may arise as a result of imperfect competition, policy restrictions, social objectives, labour and financial constraints, as well as 'lumpy' capital investment (Coelli et al., 2005).

3.2.5 The Impact of Case-Mix Adjustment and Quality on Efficiency

Accounting for differences in quality of care becomes critical when measuring efficiency. For example when measuring technical efficiency, one hospital may be using a higher quantity of inputs in comparison to another hospital in order to treat the same number of cases but the first hospital may be providing better quality care. When compared against one another, the first hospital will appear less technically efficient despite providing better quality care, which may improve the long-run efficiency of the hospital. Allocative efficiency is constrained by the required quality standard of inputs (Sherman and Zhu, 2006), which should have a direct impact on the quality of the outputs produced.

The mix of cases across hospitals will affect the measurement of both technical efficiency and allocative efficiency through differences in inputs and outputs. In order to accurately measure technical and allocative efficiency in the hospital environment, a case-mix adjustment of outputs is necessary because different hospitals do not treat a standard group of patients. Cases vary across hospitals in both type and severity. Consider two hospitals both with the same number of cases but one with many severe cases (that consume more resources) and the other with very few severe cases (that consume relatively few resources), the first hospital may require more inputs than the second hospital, as well as a different mix of inputs in order to treat patients effectively. As a result, the first hospital may appear both technically and allocatively inefficient relative to the second hospital. However, this is not a true reflection of the efficiency of the first hospital, which is in fact treating more severe cases. The efficiency scores will have been confounded by the differences in the mix of cases, capturing this difference as an inefficiency.

As a result of the relationship between technical, allocative, scale and cost efficiency, both quality of care and the mix of cases will influence the measure of cost efficiency.

3.3 Graphical Representation of the Components of Efficiency

3.3.1 Technical, Allocative and Cost Efficiency

Consider two inputs x_1 and x_2 which are used to produce a single output q. For the purposes of simplicity, it is assumed that the production process is operating under constant returns to scale. Furthermore, the graphical representations below are from an input-orientated perspective, in that input usage is proportionally reduced whilst output levels are fixed.

The unit isoquant[1], SS', in Figure 3.1 represents the various combinations of the two inputs that can be used to produce one unit of output for a fully efficient organisation (Sloman, 2006). Any organisation operating on this isoquant is producing one unit of output using the minimum level of inputs and is therefore considered to be fully technically efficient. It is therefore not possible for an organisation to operate below SS'. The isocost line, AA', represents the different combinations of inputs for a specific price (Coelli et al., 2005).

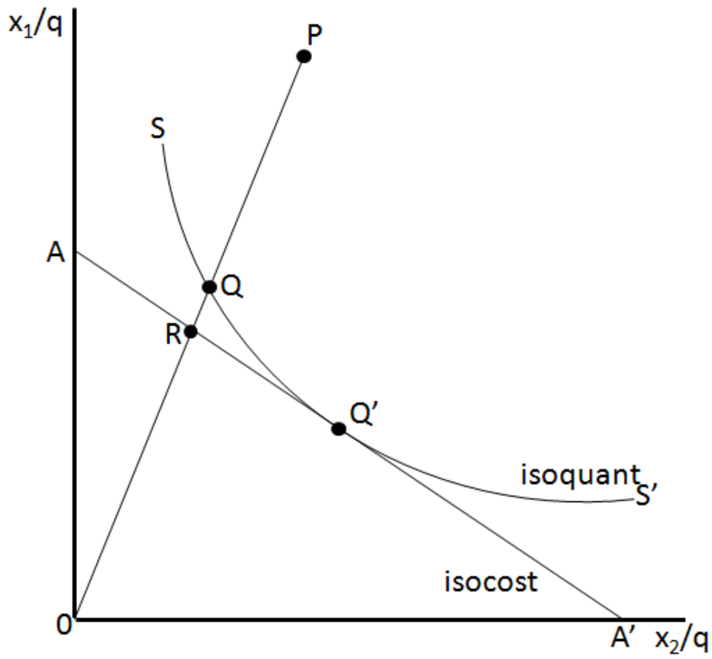

Figure 3.1: A Graphical Representation of Technical, Allocative and Cost Efficiency.
Source: Coelli et al. (2005)

Consider an organisation operating at point P. In order to operate at a technically efficient level, the organisation would have to shift down to point Q by proportionally reducing the inputs used without a reduction in output. The proportionate reduction of inputs required can be represented by the ratio $QP/0P$ (Coelli et al., 2005), this ratio is the technical inefficiency associated with the organisation. The technical efficiency (TE)

[1] A line connecting all the different combinations of inputs that can be used to produce a single output (Sloman, 2006).

of the organisation can thus be measured using the ratio

$$TE = 1 - \frac{QP}{0P} = \frac{0Q}{0P} \text{ where } 0 \leq TE \leq 1.$$

A score of one denotes a technically efficient organisation which is operating on the production frontier and that no proportionate contraction of inputs is possible. The lower the TE score, the lower the level of technical efficiency of an organisation. Technical efficiency can also be thought of as the distance that the unit lies from the efficiency frontier.

Although an organisation operating at point Q is technically efficient, it is not allocatively efficient given the relative prices of inputs, assuming substitutability of inputs. In order for the organisation to be allocatively efficient, the organisation needs to shift production from point Q to point Q' by changing the combination of inputs used to produce the single output. This is the point at which the isocost line is tangential to the isoquant. An organisation operating at point Q' will produce one unit of output for the minimum cost. The ratio $QR/0Q$ is the proportionate reduction in costs that would occur if an organisation was to operate at Q', rather than Q. Allocative efficiency (AE) can thus be measured as:

$$AE = 1 - \frac{QR}{0Q} = \frac{0R}{0Q} \text{ where } 0 \leq AE \leq 1.$$

As with technical efficiency, a score of one indicates that an organisation is allocatively efficient.

The cost efficiency of an organisation can be estimated by calculating the ratio of the minimum resource usage adjusted for price to actual observed costs (Linna, 1998; Coelli, 1996). Thus, the cost efficiency (CE) of the organisation operating at point P can be measured as the ratio of the cost efficient inputs to the inputs actually used. This definition relies on the level and mix of inputs actually used, in other words the technical and allocative efficiency of the organisation. The ratio

$$CE = \frac{0R}{0P} \text{ where } 0 \leq CE \leq 1$$

provides an indication of the level of cost efficiency of the organisation. It is important to note that it is impossible for the organisation to operate at point R. Point R represents a combination of inputs at a particular cost, however this combination is insufficient to produce one unit of output as it lies beneath the production frontier.

To summarise, an organisation is cost efficient when it is both technically efficient and allocatively efficient. Graphically, this is the point at which the isocost line is tangential to the isoquant (Coelli et al., 2005). Cost efficiency (CE) can be measured as a product of the technical efficiency, (TE), and allocative efficiency, (AE).

$$\begin{aligned} CE &= TE \times AE \\ &= \frac{0Q}{0P} \times \frac{0R}{0Q} \\ &= \frac{0R}{0P} \end{aligned}$$

3.3.2 Scale Efficiency

As with the graphical representation of technical, allocative and cost efficiencies, scale efficiency is calculated from an input-oriented perspective. Scale efficiency is straightforward in the one-input, one-output case; however it is more complex to grasp when considering multiple-input, multiple-output cases. For this reason, the graphical representation presented in this section considers only the single-input, single-output scenario.

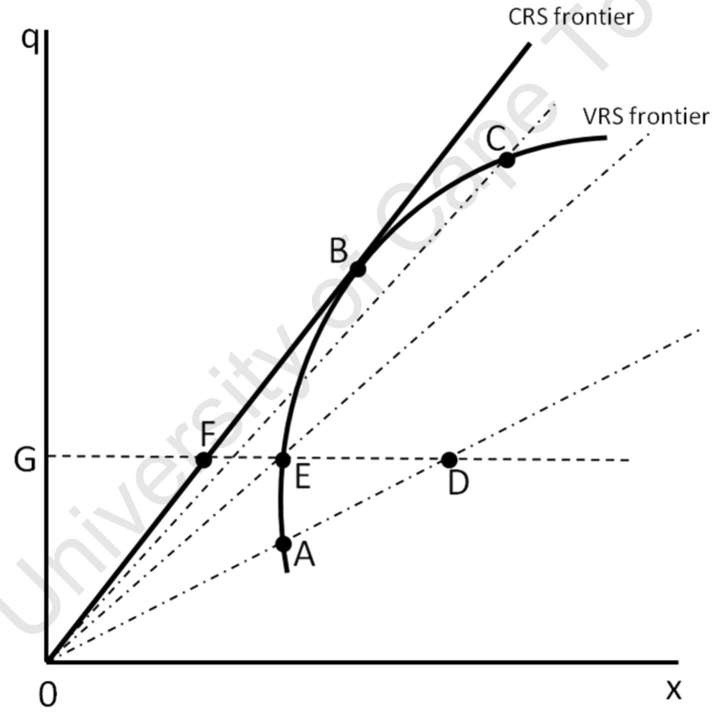

Figure 3.2: A Graphical Representation of Scale Efficiency.
Source: Coelli et al. (2005)

Figure 3.2 depicts both the constant returns to scale (CRS) production frontier and variable returns to scale (VRS) production frontier for a one-input, one-output production environment. The portion AB of the VRS frontier displays increasing returns to scale as an increase in the input, x, results in a greater than proportionate increase in the output,

q. The second portion of the VRS frontier, BC, displays decreasing returns to scale. An increase in x results in a less than proportionate increase in q.

Assuming variable returns to scale, the organisations operating at points A, B and C are all technically efficient as they are operating on the production frontier. Despite all being technically efficient, the three organisations are not equally efficient overall; this is as a result of differences in scale efficiencies (Coelli et al., 2005). A change in scale efficiency can be achieved by moving along the production frontier. For example, organisation A is operating on the increasing returns to scale frontier and could increase its scale efficiency by moving towards B by simultaneously increasing the level of the input and the output. Similarly, Organisation C could improve its scale efficiency by moving down towards B as it is operating on the decreasing returns to scale portion of the production frontier. Organisation B is said to be operating at the *most productive scale size* (MPSS) or at the *technically optimal productive scale* (TOPS) (Coelli et al., 2005). In a one-input, one-output case, it is simple to see that the MPSS point is the point at which the ratio of the output to the input is maximised. Alternatively, the MPSS point is the point at which the CRS frontier is tangential to the VRS frontier (Coelli et al., 2005).

Consider a technically inefficient organisation operating at point D. The technical efficiency score under constant returns to scale would be:

$$TE_{CRS} = \frac{GF}{GD}.$$

Under variable returns to scale the technical efficiency score would be:

$$TE_{VRS} = \frac{GE}{GD}.$$

The scale efficiency of point D can be measured by the distance of the technically efficient point E, relative to the CRS frontier (Coelli et al., 2005). Hence, scale efficiency (SE) can be measured using the following ratio:

$$SE = \frac{GF}{GE} \text{ where } 0 \leq SE \leq 1.$$

A SE score of 1 indicates that a firm is scale efficient and that no resource savings can be made through changing the scale of operations.

Very rarely is scale efficiency measured directly as shown above. Rather it is calculated as the ratio of the CRS technical efficiency score to the VRS technical efficiency score

(Coelli et al., 2005).

$$\begin{aligned} SE &= \frac{TE_{CRS}}{TE_{VRS}} \\ &= \frac{GF}{GD} \div \frac{GE}{GD} \\ &= \frac{GF}{GE} \end{aligned}$$

3.4 Efficiency and Productivity

Despite being two separate and distinct concepts, efficiency and productivity are often used interchangeably in literature (Sherman and Zhu, 2006; Coelli et al., 2005). Productivity can be defined as the ratio of an organisation's outputs to the inputs used (Sherman and Zhu, 2006), (Coelli et al., 2005) and (Caves, Christensen and Diewert, 1982).

$$productivity = \frac{outputs}{inputs}$$

In the case of a single input and a single output, the calculation of productivity is uncomplicated. However, in the case of multiple inputs and multiple outputs, calculating the productivity ratio is notably more difficult. This is as a result of outputs needing to be weighted and aggregated in a meaningful manner, rather than simply being summed. Productivity for an organisation using s inputs to produce r outputs is calculated as follows:

$$productivity = \frac{\sum_{i=1}^{r} u_i q_i}{\sum_{j=1}^{s} v_j x_j},$$

where

$$\begin{aligned} u_i &= \text{ weighting for output i} \\ q_i &= \text{ units of output i} \\ v_j &= \text{ weighting for input j} \\ x_j &= \text{ units of input j.} \end{aligned}$$

All hospitals are multiple-input, multiple-output organisations in that a range of inputs is required to produce the different services and to treat a variety of patients; hence the calculation of productivity is complex. In the hospital industry, not only is the weighting of the inputs and outputs difficult to determine, but deciding on which inputs and outputs to include is also a complex issue.

Whilst efficiency is also concerned with the ratio of inputs to outputs, it is focused on the amount of resources wasted in producing outputs, whether the wastage be associated with a sub-optimal mix of inputs, the cost of inputs or the scale of production.

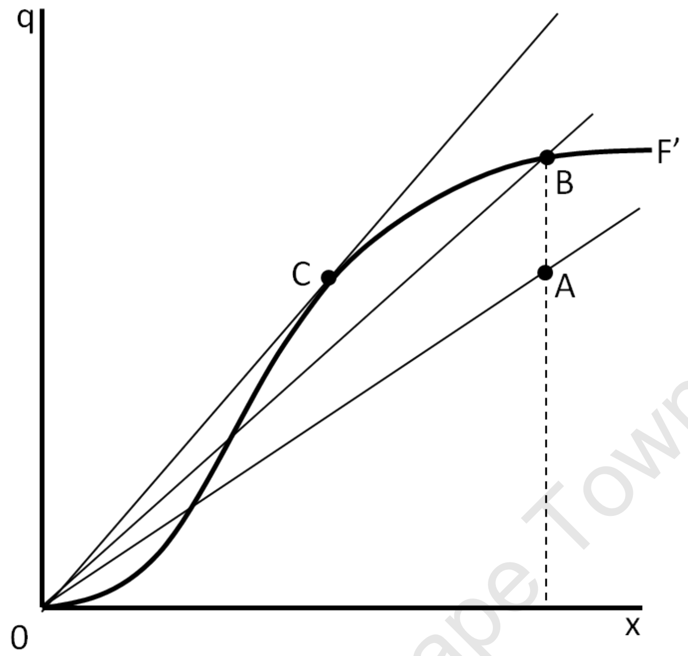

Figure 3.3: A Graphical Representation of Productivity and Efficiency.
Source: Coelli et al. (2005)

Figure 3.3 illustrates the difference between productivity and efficiency. The line $0F'$ represents the production frontier that defines the relationship between the single input, x, and the single output, q. Point A is technically inefficient as it is operating below the production frontier, whilst points B and C are technically efficient. The slopes, q/x, of the rays $0A$, $0B$ and $0C$ through the origin, offer an indication of the level of productivity. The steeper the slope, the higher the level of productivity. Whilst points B and C are technically efficient, they are not equally productive.

Consider an organisation operating at point A. If the organisation were to shift to point B, it would become technically efficient by operating on the production fronter and it would become more productive by increasing the ratio of the output to the input. If an organisation operating at point B were to move to point C, there would be an increase in scale efficiency as it would move to the MPSS point, the point at which the productivity of the organisation is maximised. This shift would increase the productivity of the organisation as the ratio of the output to the input increases. It is clear from Figure 3.3 that a change in productivity may give rise to a change in one of the components of efficiency, but productivity and efficiency do not strictly have the same definition nor the same measurement. In a sense, productivity encompasses efficiency.

Another aspect of productivity is the change in the set of feasible combinations of inputs to produce a given set of inputs over time. This change results in either the expansion or the contraction of the production frontier over different periods (Coelli et al., 2005; Balk, 2001; Caves et al., 1982). This is known as technological change.

Technological change and efficiency change are two independent elements of productivity. These two elements can occur independently of one another; however, it is more often a combination of the two changes which gives rise to a change in productivity (Balk, 2001).

3.5 Measuring Efficiency

3.5.1 An Introduction to Measuring Efficiency

The majority of efficiency measurement techniques estimate the cost or production frontier in order to derive efficiency scores. The frontier is established by analysing the organisations that produce the highest level of output given a fixed level of input, or alternatively produce a given level of output using the smallest quantities of input (Hollingsworth, 2003). Hollingsworth et al. (1999) note that there are two major features which distinguish different frontier estimation methods. Firstly, whether the method is parametric or non-parametric and secondly, whether the method is stochastic or deterministic. Table 3.1 summarises these different methods.

Table 3.1: Summary of efficiency measurement methods.
Source: Hollingsworth et al. (1999).

	Parametric	Non-parametric
Deterministic	Parametric mathematical programming and Deterministic (econometric) frontier analysis	Data envelopment analysis
Stochastic	Stochastic (econometric) frontier analysis	Stochastic data envelopment analysis

There are two techniques in the hospital efficiency literature which are most commonly used to measure efficiency: Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA). Data Envelopment Analysis is a non-parametric linear programming technique which, according to O'Neill, Rauner, Heidenberger and Kraus (2008) and Worthington (2004), is the most commonly used non-parametric efficiency estimation method in hospital efficiency literature. SFA is similar to DEA in that it constructs a production frontier, however it is a parametric and stochastic method. DEA is discussed in more detail in Section 3.5.2 and SFA in Section 3.5.3.

3.5.2 Data Envelopment Analysis

DEA is a non-parametric, multiple input – multiple output linear programming technique. Based on the multiple input and multiple output data, a non-parametric piece-wise frontier is derived by explicitly considering the relative mix of inputs and outputs of each unit (Sherman and Zhu, 2006). The best practice units, those organisations that appear to be the most efficient relative to all other organisations, are those that are producing the maximum level of output for a given input level or alternatively those that are producing a given level of output using the minimum required inputs. The constructed frontier is defined by these identified best practice units (Coelli et al., 2005). The efficiency of all the other organisations in the sample is measured relative to these best practice units (Hollingsworth et al., 1999). Furthermore, comparing the best practice units for one time period relative to the best practice units in the previous time period determines the movement of the frontier over time (the technological change) (Hollingsworth et al., 1999).

The efficiency scores derived using DEA are sensitive to the number of inputs and outputs used, as well as the number of organisations in the sample being analysed (Nguyen and Coelli, 2009). The higher the number of inputs and outputs used, all other things being equal, the more difficult it becomes to find comparable organisations because the combinations of inputs used to produce outputs become more heterogeneous (Nguyen and Coelli, 2009). If no comparable organisations are found, then an organisation is classified as efficient (Sherman and Zhu, 2006). In this way, DEA is said to give organisations the benefit of the doubt by assuming they are fully efficient in the presence of no further information. A similar problem arises when there are too few organisations being compared to one another. If there are very few organisations and they are not comparable, efficiency scores will be inflated (Sherman and Zhu, 2006).

DEA suffers from a number of drawbacks. DEA does not allow for a random error term, as in SFA (Jacobs, 2001). As a result DEA may overstate estimated inefficiencies as all random noise is classified as inefficiency (Jacobs, 2001). Many have argued that this is a major shortcoming of DEA. This deficiency can be managed to some extent by using annual data, which reduces the effect of seasonal changes and irregular observations which may be present in daily data, and by increasing the volume of data. Sherman and Zhu (2006) and Worthington (2004) warn that the efficiency scores derived using DEA are *relative* efficiency scores and give no indication about the *absolute* efficiency about individual organisations. As a result, this method should not be used to analyse incomplete or biased samples of a group in order to draw conclusions about entire group.

A third disadvantage, in comparison to other efficiency measurement methods, is that derived efficiency scores may be highly sensitive to data errors (Jacobs, 2001). Although data errors do not exclusively impact efficiency scores calculated using DEA, they also impact efficiency scores calculated using other methods such as SFA, the DEA calculated efficiency scores behave slightly differently. Instead of only affecting the score for the one hospital, a data error for an input for one hospital will impact the efficiency scores calculated for **all** other comparable hospitals. This is because efficiency scores are calculated relative to other hospitals and a change in one efficiency score will result in a change to others (Jacobs, 2001).

Despite theses practical limitations, DEA is a popular methodology when analysing hospital efficiency (O'Neill et al., 2008). DEA has a number of advantages over other methods such as SFA. Unlike parametric methods, no assumption is required regarding the relationship between inputs and outputs (O'Neill et al., 2008), (Jacobs, 2001), and (Charnes, Cooper and Rhodes, 1978). This is considered to be a major advantage because it reduces specification error. No assumption is required about the distribution of the production function. Rather, the production function is inferred from the available data. Furthermore, no prior information is required regarding the relative weights of inputs and outputs, which is important as this data is often unknown or, if known, not readily available (O'Neill et al., 2008). Another advantage of DEA is that inputs can be measured in their natural units and do not have to be transformed in any manner (Charnes et al., 1978).

Stochastic DEA (SDEA) is identical to DEA in all aspects other than that of the specification of inputs and outputs. Rather than specifying inputs and outputs as constants, under an SDEA model these are stochastic in nature (Kao and Liu, 2009). SDEA has not been widely used in any hospital efficiency analyses to date.

3.5.3 Stochastic Frontier Analysis

SFA is a parametric alternative to DEA which is used in the estimation of production frontiers. As the name suggests, SFA is a stochastic technique and thus allows for random noise in the estimation of the production frontier and inefficiency terms. SFA divides the stochastic error term into an error term, which accounts for both random noise and systemic errors, and an inefficiency component (Nguyen and Coelli, 2009). For this reason, it is arguably preferable to other parametric methods. This method is used to construct a smooth parametric frontier, rather than a piece-wise linear-segmented frontier which is constructed using DEA (Jacobs, 2001). In order to construct this smooth frontier, assumptions are required about both the functional form of the relationship between the inputs and outputs, also known as the technology, as well as the distribution of the inefficiency term, the second component of the error term.

The most common functions used to define the relationship between the inputs and the outputs are the Cobb-Douglas and the Trans-log functions (Nguyen and Coelli, 2009; Worthington, 2004). The Cobb-Douglas function is a popular production function in economics and requires only a few parameters to be specified (Nguyen and Coelli, 2009). The use of the Trans-log function, which is simply a generalisation of the Cobb-Douglas function, allows for a second order approximation of an arbitrary functional form for the technology (Nguyen and Coelli, 2009). Jacobs (2001) notes that the need to specify the relationship between inputs and outputs may prove to be too restrictive in complex production environments with multiple inputs and multiple outputs, such as hospitals.

The distribution of the inefficiency term, the second distributional assumption required, is often specified as the half-normal, truncated normal, gamma or exponential distribution (Nguyen and Coelli, 2009). Whilst different distributions for the inefficiency term may give rise to different efficiency estimates, Nguyen and Coelli (2009) note that the impact of using different distributions is not well documented in the literature. A lack of research into determining which distribution to use in which scenarios and contexts has resulted in many researchers and economists choosing distributions based on computational convenience, rather than appropriateness (Nguyen and Coelli, 2009).

Having chosen the distributional forms of the production frontier and the inefficiency terms, the parameters of these distributions need to be estimated. In the case of single-output technologies, parameters are often approximated using maximum likelihood estimation (Coelli et al., 2005). When approximating the parameters of multiple-output technologies, cost frontiers are used if it can be assumed that the organisations being analysed aim to maximise profits and both input and output prices are available (Coelli et al., 2005). If it is inappropriate to assume that organisations maximise profits or if the required data are unavailable, distance functions are often used to estimate the parameter values (Coelli et al., 2005).

Despite the calculation of inefficiency terms being highly sensitive to the specification of the distributional forms of technology and inefficiency, in previous studies it has been shown that it is not highly sensitive to data errors or data changes at the individual organisation level (Jacobs, 2001). This is likely to be as a result of the efficiency scores being estimated from distributions which are parameterised using all available data and not just the data of a single organisation. For this reason, SFA may be well suited to analysing the efficiency of South African public hospitals where the availability and quality of data is poor.

The key advantage of SFA, as a parametric method over non-parametric methods, is the ability of the method to account for random noise such that this noise does not confound inefficiency scores (Coelli et al., 2005). However, SFA is sensitive to the specification of the distributional form of both the inefficiency term and production frontier.

The extent to which SFA is superior to non-parametric methods, in terms of separating out inefficiency and random noise, will depend on the appropriateness of the assumption of the distributional form of the inefficiency terms and the production frontier (Jacobs, 2001).

SFA has been used in a number of hospital efficiency studies around the world. O'Neill et al. (2008), Worthington (2004) and Hollingsworth (2003) all provide good overviews of hospital efficiency studies that have used SFA, as well as the results of those studies. To date, there is only one study, by Kinfu (2011), that has used SFA to analyse efficiency in the health care sector in South Africa.

3.5.4 Selecting an Efficiency Measurement Method

Based on the advantages and disadvantages of different efficiency measurement techniques, outlined in Sections 3.5.2 and 3.5.3, DEA was chosen as the preferred technique for this hospital efficiency analysis. There are four major reasons for selecting DEA. Firstly, DEA can more easily handle multiple inputs and multiple outputs, which is appropriate, given the nature of hospital production processes. Secondly, the flexibility of DEA allows different methods of case-mix adjustments to be analysed, which is central to this dissertation. Thirdly, DEA requires no *a priori* assumptions regarding the distribution of the production frontier and the efficiency term. Although Kinfu (2011) analyses hospital efficiency in South Africa using SFA, beyond this study there is no evidence to confirm which distributional assumptions are appropriate for the South African private hospital industry. For the same reason, it was decided that DEA was preferable to SDEA. Finally, the data in the private hospital sector in South Africa is of good quality which reduces the concerns raised about the impact of incorrect data on efficiency scores when using DEA.

A detailed description of DEA is provided in Chapter 4.

Chapter 4

An In-Depth Look at Data Envelopment Analysis

4.1 Overview

This section provides an overview of the key components of constructing a DEA model; namely, the selection of inputs and outputs, the returns to scale, the orientation of the model, the mathematical construction of a model, the inclusion of panel data, the use of DEA as a management tool and the limitations of DEA models.

4.2 Inputs and Outputs

The selection of inputs and outputs should be guided by the range of resources required to provide a set of outputs (Sherman, 1984); however, the choice of inputs and outputs is often limited by the quality of data available. The inputs and outputs selected should all be relevant to the production process and should be sufficient to capture the nature and complexity of the process (Nguyen and Coelli, 2009). A number of authors provide comprehensive overviews of the inputs and outputs most often used in hospital efficiency studies (Nguyen and Coelli, 2009; O'Neill et al., 2008; Worthington, 2004; Hollingsworth et al., 1999), these inputs and outputs are discussed in detail below. This is followed by a discussion of the number of variables used, as the number of variables used and estimated efficiency scores are positively correlated (Nguyen and Coelli, 2009). Furthermore, the number of organisations (known as DMUs in the DEA environment) being evaluated will also influence the efficiency scores. Nguyen and Coelli (2009) raise two major concerns with respect to variable selection, namely, incorrect aggregation of variables and omission of relevant variables. The impact of the number of variables and the number of organisations used, as well as the final two concerns raised, are discussed in detail below.

Both Worthington (2004) and Hollingsworth et al. (1999) report that the majority of studies using DEA to evaluate hospital efficiency use two major input categories, namely, labour and capital. Nguyen and Coelli (2009) identify goods as a third input category. Labour can be measured in one of two ways, either in monetary terms such as salaries and wages, or in other quantitative terms such as numbers of staff or hours worked. O'Neill et al. (2008) suggest that using the number of staff is a more appropriate measure because of differences in salaries and wages across both medical institutions and geographical locations. This is problematic as inputs are assumed to be identical across units and a higher salary figure should result in a higher output level. In reality, this is not necessarily the case because differences in salaries and wages across institutions and locations will not necessarily result in a higher output.

The ideal measure of capital is the flow of capital services[1] (Worthington, 2004). However, many studies use an estimation of capital stock, rather than attributing capital use to the period in question. This is largely as a result of the difficulties and complexity of measuring the existing infrastructure and measuring the utilisation of capital goods over the period (Nguyen and Coelli, 2009). As a result, in most hospital efficiency studies, a proxy for capital is to be used. Worthington (2004) warns that using a proxy that estimates capital stock, rather than capital consumed, can result in a misspecification and overestimation of the use of capital, which will bias efficiency score estimates. The number of staffed beds, or total number of beds, is often used as a proxy for capital because of the high correlation between the number of beds and the size of the hospital and, hence, capital investment (Clement, Vladmanis, Bazzoli, Zhao and Chukmaitov, 2008). The third input commonly used in hospital efficiency studies, namely goods, is often defined as the cost of medical goods, non-medical goods and materials (Nguyen and Coelli, 2009).

The choice of outputs is more complex than that of inputs. The conceptual output would be an improvement in health status of a patient, such as increased longevity, or decreased incidence of disease, or improved quality of life (Nguyen and Coelli, 2009; Worthington, 2004). Theoretically, these conceptual outputs, known as final health outcomes, can be measured by determining the difference between a patient's health status after having completed the required treatment and the health status of the patient had he or she not undergone any treatment (Nguyen and Coelli, 2009). Worthington (2004) notes that it is difficult, if not impossible, to measure this change in health status. As a result, in most hospital efficiency studies, outputs which are related to an improvement in the health status of patients have to be used as a proxy for the conceptual output (Nguyen and Coelli, 2009). Examples of such proxies are the number of surgeries, the number of emergency visits, the number of inpatient days for acute care and intensive care, and the

[1]This is the capital consumed in order to produce the outputs (Nguyen and Coelli, 2009).

number of outpatient cases (O'Neill et al., 2008; Worthington, 2004). These proxies are often referred to as intermediate health outcomes. Chilingerian (1994) suggests that the use of intermediate health outcomes is acceptable, as long as these outputs are adjusted for differences in the complexity of cases, as well as differences in the severity of cases. It is these adjustments that are the key focus of this dissertation. Alternatively, the use of intermediate health outcomes is appropriate when there is sufficient evidence to suggest that these intermediate activities do lead to an improvement in health status (Nguyen and Coelli, 2009). If this is not the case, these proxies may be less reliable as a measure of final health outcomes and efficiency scores may be biased.

Having identified the relevant inputs and outputs, an important consideration is the total number of variables used. The total number of variables used impacts the power of the DEA model, that is the ability of the model to identify inefficient units. In general, as the number of inputs and outputs included in the model increases, all other things being equal, it is harder to identify comparable units using a similar mix of inputs to produce a particular mix of outputs (Sherman and Zhu, 2006). In the event that there are no comparable units, a unit is classified as fully efficient and average efficiency scores increase. However, the magnitude of the effect of adding a variable depends on the correlation between the new variable and the existing variables (Nguyen and Coelli, 2009). For example, if a variable is added that is highly correlated with existing variables, adding this variable is unlikely to significantly affect the results. However, if the added variable is not strongly correlated with the existing variables, the impact of including this variable on mean efficiency scores can be remarkable. There are two common mistakes made when trying to reduce the number of variables in a model, namely omission of important variables and inappropriate aggregation of variables (Nguyen and Coelli, 2009).

The omission of significant variables can bias efficiency scores. Omitting variables usually occurs as a result of missing data or imperfect measures for variables (Nguyen and Coelli, 2009). A common missing input variable in hospital efficiency studies is a measure of capital. On the output side, many studies omit teaching and research variables (Nguyen and Coelli, 2009). Prior South African public sector hospital efficiency studies omitted variables, such as adjustments for case mix, due to poor data availability and quality.

Aggregation of variables usually occurs as a result of zero values in variables, missing data and a constraint on the number of variables included in the final model (Nguyen and Coelli, 2009). The inappropriate aggregation of variables can bias efficiency scores. In hospital efficiency studies, doctors and nurses are usually aggregated to create a 'labour' input (Nguyen and Coelli, 2009). The aggregate may be weighted or unweighted. On the output side, the number of surgical cases, day cases, ambulatory cases and maternity cases are often aggregated into one variable representing all cases (Nguyen and Coelli, 2009).

Early studies used an unweighted aggregation of cases, whilst later studies have used case-mix information to weight cases according to resource consumption (Nguyen and Coelli, 2009). Nguyen and Coelli (2009) warn that an unweighted aggregated cases variable may bias results when particular units which treat more complex cases are compared against units treating less complex cases. A basic example of aggregation bias is illustrated in Figure 4.1.

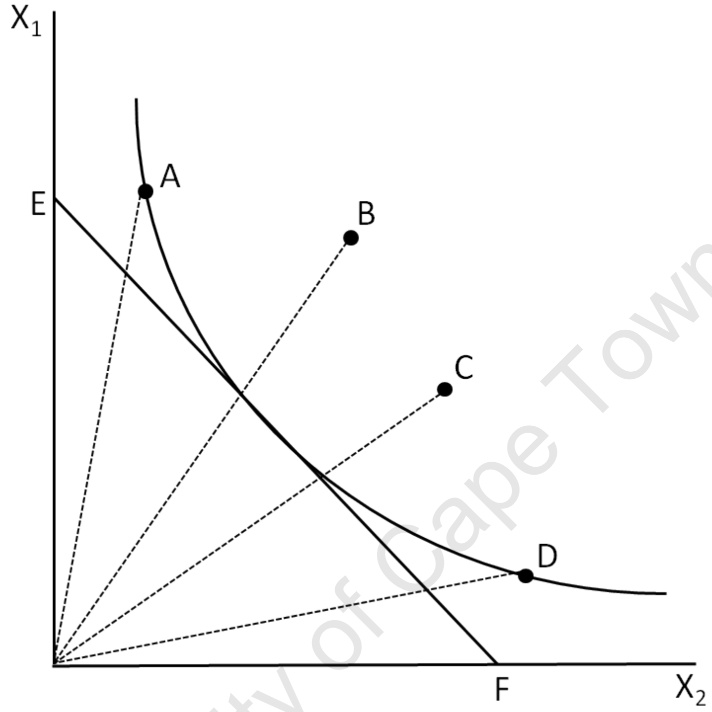

Figure 4.1: The impact of inappropriate aggregation of inputs on efficiency scores.
Source: Nguyen and Coelli (2009).

The curve, AD, is the convex isoquant, whilst the 45 degree straight line, EF, is a linear aggregation of inputs X_1 and X_2 of equal weights. Under the convex isoquant, both A and D are fully efficient whilst B and C are inefficient. However, under the linear isoquant, EF, all four units appear to be inefficient. In this case, the aggregation will result in a decrease in the mean efficiency scores in comparison to the mean efficiency scores under the convex isoquant. Barnum and Gleason (2012) reported similar findings in that linear aggregation of the same type of input, such as a linear aggregation of nurses and doctors into a single labour input, creates a downward bias on technical efficiency scores.

It is important to note that whilst aggregation may cause biases in the measurement of efficiency, aggregation is preferable to omitting significant variables (Nguyen and Coelli, 2009). Despite aggregation being problematic, all inputs and outputs are all still captured in the production process to some degree, unlike the omission of variables. Aggregation bias can be managed by determining appropriate weights for the relative inputs and outputs, or alternatively by using non-linear aggregation techniques such as the Fisher index number formula (Nguyen and Coelli, 2009).

4.3 Sample Size

The number of units analysed impacts efficiency scores. All other things being equal, increasing the sample size may push up the production frontier and decrease the mean efficiency scores (Nguyen and Coelli, 2009). This is particularly true for a small sample. Consider a small sample with few comparable units; most units will be classified as fully efficient in the light of no further evidence. If more units are added to the sample, this increases the number of comparable units and efficiency scores are likely to decrease as a result. This is illustrated in Figure 4.2.

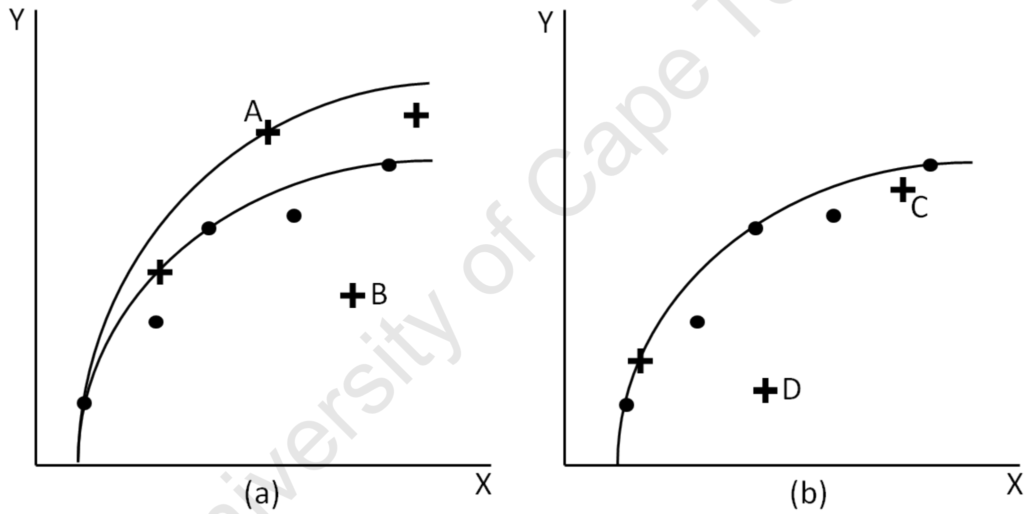

Figure 4.2: The impact of increasing the sample size.
Source: Nguyen and Coelli (2009).

Consider graphs (a) and (b) in Figure 4.2. There are two types of observations; the first being existing units represented by the dots, and the second being the new units represented by the crosses. In graph (a), the new units form a new frontier and those units that were classified as efficient are no longer necessarily still efficient. However, the new units in graph (b) fall within the existing frontier and there is no change to the production frontier. A new unit in the sample which does not change the frontier, will not change the status (whether a unit is efficient or not) of the existing units. Nguyen and Coelli (2009) note that when the sample size is large, the addition of units has very little impact on mean efficiency. It is important to note that the impact of additional units does depend on how homogeneous the existing group of units is as it impacts the likelihood of finding a comparable unit.

4.4 Returns to Scale

The returns to scale refers to the rate of substitution between inputs and outputs (Sherman and Zhu, 2006). A DEA model can be specified as either having constant returns to scale (CRS) or variable returns to scale (VRS). VRS incorporates both increasing returns to scale (IRS) and decreasing returns to scale (DRS). CRS implies that a proportionate increase in all inputs will result in an equally proportionate increase in outputs (Coelli et al., 2005). IRS occurs when a proportionate increase in all the inputs leads to a more than proportionate increase in outputs (Coelli et al., 2005). Similarly, DRS occurs when an increase in all the inputs leads to a less than proportionate increase in outputs. The assumption of CRS is appropriate when all firms are operating at the optimal size (Coelli et al., 2005).

If the model is specified as CRS when not all the units are operating at the optimal size, technical efficiency scores will be biased (Nguyen and Coelli, 2009). Diseconomies of scale are classified as inefficiencies when a CRS assumption is used in a DEA model but the inefficiencies may be a result of the size of the unit or volume of production, rather than the excess use of resources (Sherman and Zhu, 2006). For example, if a large and a small unit are compared against one another, the small unit may appear inefficient as it produces a quarter of the output of the large unit but uses half of the inputs. A CRS DEA model will classify the small unit as inefficient. However, if the unit is operating under IRS, the inefficiency may be exclusively due to the size of the unit. This confounding of efficiency scores is illustrated in Figure 4.3.

Figure 4.3 illustrates a one input, one output production model. The ray OR represents the CRS production frontier and the curve CRD represents the VRS production frontier. Point R is technically efficient on the CRS frontier, whilst points C, R and D are all efficient on the VRS frontier. If the model was specified as CRS, points C and D would be classified as inefficient. Consider point P which is technically inefficient under both CRS and VRS assumptions: under CRS, the input-oriented technical inefficiency of point P is PP_C/AP; similarly, under VRS, the input-oriented technical inefficiency of point P is PP_V/AP. From this graph, it is clear that imposing a CRS assumption decreases the efficiency scores. Coelli et al. (2005) note that the VRS technical efficiency scores will be greater than, or equal to, the efficiency scores estimated when assuming CRS. Furthermore, if there is a difference between CRS and VRS efficiency scores, scale inefficiencies exist (Coelli et al., 2005).

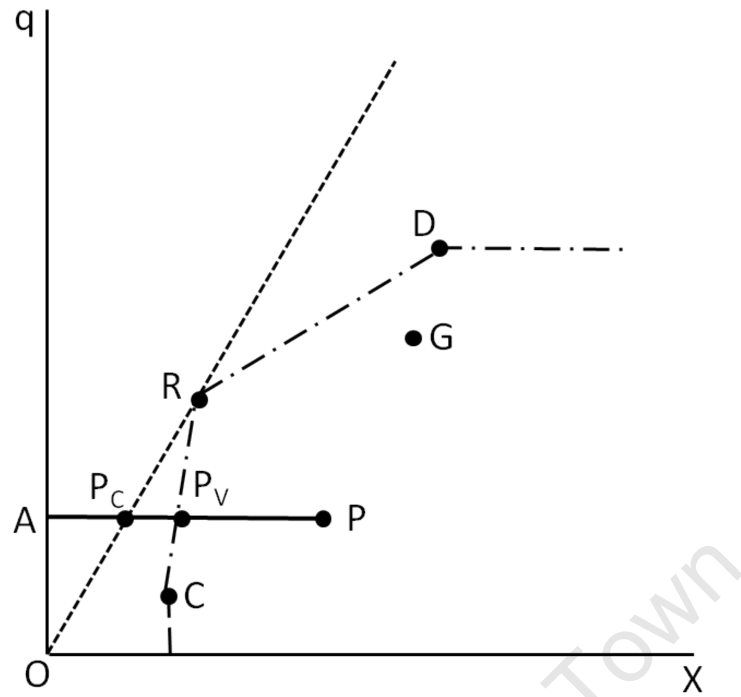

Figure 4.3: The impact of the returns to scale assumption.
Source: Nguyen and Coelli (2009).

In the analysis of hospital efficiency, numerous studies have established that scale economies are present in the hospital environment (O'Neill et al., 2008). Despite this evidence, many studies have used a CRS assumption. In order to avoid confounding efficiency scores, Sherman and Zhu (2006) suggest that a VRS model be used in conjunction with a CRS model in order to analyse whether scale efficiencies are present.

4.5 Orientation

Another methodological consideration is the orientation of the DEA model. Sherman and Zhu (2006) define the two different orientations of DEA models:

Input-Oriented DEA Models optimise the use of inputs, whilst holding outputs constant. This is most often used when the demand for outputs cannot be controlled, but management is able to influence or control resource usage.

Output-Oriented DEA Models maximise output volume and mix for a given a level of inputs. This orientation is often used when units are faced with input constraints.

Nguyen and Coelli (2009) suggest that the choice of orientation should be driven by the objectives of the analysis and any management constraints. In the health care environment, most studies use an input-oriented model as the focus is predominantly on reducing and controlling costs, rather than increasing the demand for health care (O'Neill et al., 2008). For example, a hospital is able to monitor and manage the resources used to treat patients, but it cannot influence the number of patients requiring treatment in the emergency room or the number of non-elective surgeries (Sherman and Zhu, 2006).

The same frontier is estimated when using either an input- or output-orientation, therefore the same set of firms are identified as efficient under both orientations (Coelli et al., 2005). However, the two orientations may produce slightly different efficiency measures for the inefficient firms. More specifically, if CRS is assumed, the efficiency measures will be equal under both orientations; however, under VRS the efficiency scores will differ (Coelli et al., 2005).

4.6 Mathematical Formulation of DEA

4.6.1 An Introduction to the Models

In 1978, Charnes, Cooper and Rhodes published an influential paper that provided a new definition of efficiency, using a linear programming methodology which has become widely used. This model became known as the *CCR model*. This model was an input-orientated model which assumed constant returns to scale. Over the years it has been adapted and extended.

This section summarises the two basic input-orientated linear programming models, the multiplier model and the envelopment model. It also offers a brief overview on models which include slacks, and analyses how returns to scale can be incorporated into a model. The final section summarises all the models. For a detailed description of output-orientated models, please refer to Sherman and Zhu (2006).

4.6.2 The CCR Model

Under the standard CCR model, the ratio of outputs to inputs is maximised; effectively the productivity of the organisation (referred to as a decision-making unit (DMU) in DEA literature) is maximised. A linear programming method is used to determine a set of weights for the outputs and inputs from the data, u and v respectively, in order to maximise the efficiency score, θ, for the DMU being evaluated (Sherman and Zhu, 2006). An advantage of this model is that no *a priori* information is needed regarding these weights. The weights are determined objectively and are optimal weights such that any other combination of weights will make the DMU appear either equally efficient or less efficient.

Each DMU is made to look as efficient as possible, given the inputs used and outputs produced. Note that an organisation which is efficient will not be identified as inefficient by the model (Sherman and Zhu, 2006). For these reasons, DEA is said to give DMUs the benefit of the doubt (Sherman and Zhu, 2006). The inefficiency scores obtained, calculated as $(1 - \theta)$, tend to understate actual inefficiencies, and any inefficiencies which are identified are real and adjustments of inputs and outputs can be made to improve the efficiency of the DMU (Sherman and Zhu, 2006). As all inefficiencies identified are real and the inefficiency scores err on the side of caution, this method can be used with confidence.

The discussion regarding the mathematical formulation of DEA begins with the description of the basic CCR model in ratio form. A standard linear programme, well-described in Sherman and Zhu (2006), is used to specify the model.

Consider the variables:

$$\begin{aligned}
n &= \text{number of DMUs being compared} \\
\theta &= \text{efficiency score} \\
s &= \text{number of outputs} \\
m &= \text{number of inputs} \\
y_{rj} &= \text{amount of output } r \text{ produced by DMU } j \\
x_{ij} &= \text{amount of input } i \text{ used by DMU } j \\
u_r &= \text{the weight assigned to output } r \\
v_i &= \text{the weight assigned to input } i
\end{aligned}$$

The efficiency score is maximised:

$$max\ \theta_j \tag{4.1}$$

$$\text{where } \theta = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}$$

This is subject to the constraint that a DMU will not be more than 100% efficient:

$$\begin{aligned}
\frac{\sum_{r=1}^{s} u_r y_{rn}}{\sum_{i=1}^{m} v_i x_{in}} &\leq 1 \text{ for all } n; \text{and} \\
u_1, ..., u_s &\geq 0; \text{ and} \\
v_1, ..., v_m &\geq 0.
\end{aligned}$$

This model is run separately for all n. The model produces three sets of outputs for each DMU; namely, an efficiency score, θ; a matrix of input weights, v; and a matrix of output weights, u. Each DMU is assigned a set of weights that are the most favourable and that maximise the efficiency score (Coelli et al., 2005). Therefore, each DMU will have a different set of optimal weights.

The efficiency score indicates how efficient a DMU is, in comparison to the other units in the sample. A fully efficient DMU has a score of one and lies on the efficiency frontier. A DMU with a score of less than one is inefficient, according to the Farrell (1957) definition of efficiency. The quantity of input savings possible can be calculated bu multiplying the inefficiency score, $(1-\theta)$, and the value of each input.

The drawback of this basic CCR model is that there are multiple solutions available (Coelli et al., 2005). For example, if $(u*, v*)$ was a solution to this linear programming problem, then $(\alpha u*, \alpha v*)$ would also be a solution. In order to obtain only one solution, an additional constraint is required. This model, known as the multiplier model, is defined in Sherman and Zhu (2006) as follows:

$$max\ \theta_1 \tag{4.2}$$

$$\text{where } \theta = \sum_{r=1}^{s} u_r y_{r1}$$

such that:

$$\sum_{i=1}^{m} v_i x_{i1} = 1;\ \text{and}$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \text{ for } j = 1, ..., n;\ \text{and}$$

$$u_r, v_i \geq 0.$$

The constraint that limits the weighted sum of the inputs to one ensures that there is only one solution to the linear programming problem. The efficiency score of a DMU is maximised by the chosen weights such that any other combination of weights will result in the DMU being equally or less efficient (Charnes et al., 1978).

The dual of the multiplier model is known as the envelopment model. Although they take different forms, the two models are equivalent and produce identical solutions. The standard envelopment model is defined by Sherman and Zhu (2006) as follows:

$$min\ \theta_1 \tag{4.3}$$

subject to:

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{i1} \text{ for } i = 1, 2, \ldots, m; \text{ and}$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r1} \text{ for } r = 1, 2, \ldots, s; \text{ and}$$
$$\lambda_j \geq 0 \text{ for } j = 1, 2, \ldots, n.$$

where

$$n = \text{number of units being compared}$$
$$\theta = \text{efficiency score}$$
$$s = \text{number of outputs}$$
$$m = \text{number of inputs}$$
$$y_{rj} = \text{amount of output } r \text{ used by DMU } j$$
$$x_{ij} = \text{amount of input } i \text{ used by DMU } j$$
$$\lambda_j = \text{the weight associated with organisation } j$$

This linear programming problem aims to minimise θ under three constraints:

1. For each input i, the weighted sum of input i across all the other DMUs, is less than or equal to input i of the DMU being analysed multiplied by the efficiency score θ.
2. Similarly, for each output r, the weighted sum of outputs j across all the other DMUs, is greater than or equal to output r of the DMU being analysed.
3. The weights, λ, should all be greater than or equal to zero.

Rather than weight the individual inputs and outputs, as in the multiplier model, the envelopment model assigns a weight, λ, to each DMU. Using this model, the efficiency reference set (ERS) for each DMU can be obtained; that is, when minimising θ, those DMUs with non-zero λ values are the DMUs in the ERS. The ERS is a group of comparable hospitals, also known as peers. DMUs that are part of an ERS, or that have non-zero λ values, are fully efficient (Sherman and Zhu, 2006). When θ is below one for a particular DMU, the DMUs that form part of the ERS will produce as much or more output than the the DMU being analysed, for a given level of input (Sherman and Zhu, 2006). In the case where no λ values can be found to minimise θ, or reduce it to less than one, this DMU will be allocated an efficiency score of one as there is no visible opportunity to improve efficiency in comparison to the other DMUs (Sherman and Zhu, 2006).

The interpretation of the envelopment model is more intuitive than that of the multiplier model. The aim of the linear programming problem is to minimise the efficiency score, θ, for each DMU. This minimisation is achieved by radially contracting the input set, x, as far as possible, whilst still remaining within the feasible input set (Coelli et al., 2005). Whilst inputs are radially contracted, non-negative values are simultaneously assigned to the weights, λ. These weights are used to project a point on to the surface of the efficiency frontier (Coelli et al., 2005). This point is constructed as a weighted sum of the inputs and outputs of each firm in the ERS, using the respective λ values (Coelli et al., 2005). In matrix notation, the projected point on the efficiency frontier can be written as $(\lambda\mathbf{X}, \lambda\mathbf{Y})$ (Coelli et al., 2005).

Although the two models produce identical results, the envelopment model is most often used (Coelli, 1996). The reason for this is that there are fewer constraints to solve when using the envelopment model. The multiplier model has a constraint for each DMU of the sample, as well as the constraint that the weighted sum of the inputs is one, so there are $n + 1$ constraints. The envelopment model has a constraint for each input and output used, so there are only $n + m$ constraints.

Models that incorporate slacks are discussed in the following section.

4.6.3 Incorporating Slacks

It is important to note that when using the multiplier and envelopment model, there is the possibility that inputs can be reduced further, or alternatively outputs could be proportionately increased further, without impacting the efficiency score of the DMU and without violating any of the constraints of the model (Charnes et al., 1978). These are known as input and output slacks respectively.

Slacks arise as a result of the piece-wise linear non-parametric frontier that is constructed using DEA. If sections of the frontier run parallel to the axes, a reduction of inputs may be possible without altering the efficiency score (Coelli et al., 2005). For example, consider the case of 2 inputs (x_1 and x_2)and one output (q), where the x-axis is x_1/q and the y-axis is x_2/q, and part of the production frontier runs parallel to the y-axis. This would mean that any point sitting on this parallel section would be able to reduce the usage of input x_2 but still maintain the same level of output.

Sherman and Zhu (2006) recommend solving the following linear programming problem, after solving either the multiplier model (Equation 4.2) or the envelopment model (Equation 4.3), in order to ascertain the size of the slacks associated with each DMU.

$$max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \tag{4.4}$$

subject to:

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{i0} \;\; i = 1, 2, ..., m;$$
$$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{r0} \;\; r = 1, 2, ..., s; \text{and}$$
$$\lambda_j \geq 0 \;\; j = 1, 2, ..., n;$$

where θ^* is the DEA efficiency score obtained from the above equations 4.2 and 4.3, s_i^- is the input slack, and s_r^+ is the output slack.

Essentially, calculating efficiency scores becomes a two stage process. The first step is to calculate θ^* using Equations 4.2 or 4.3, whilst ignoring the slacks. The second step is then to optimise the slacks in Equation 4.4.

This second stage linear programming model (Equation 4.4), as with both the multiplier and the envelopment models, must be solved for all n DMUs. The input and output slacks for each DMU are maximised subject to the constraint that, for each input i, the weighted sum of the inputs across DMUs, plus the slack should be equal to the optimal quantity of input that should be used for the DMU being analysed. Furthermore, slacks are maximised subject to, for each output r, the weighted sum of the outputs across DMUs, less the output slack being equal to the actual output produced by the DMU being analysed.

The slacks obtained in Equation 4.4 can be used to move an inefficient DMU onto the efficient frontier using the following formulae (Sherman and Zhu, 2006):

$$\hat{x}_{i0} = \theta^* x_{i0} - s_i^{-*} \; i = 1, 2, ..., m \tag{4.5}$$

$$\hat{y}_{r0} = y_{r0} + s_r^{+*} \; r = 1, 2, ..., s \tag{4.6}$$

where s_i^{-*} and s_r^{+*} are from Equation 4.4, and θ^* is obtained from Equation 4.2 or 4.3.

There are two major problems associated with this two-stage slack linear programming problem. Firstly, slacks are maximised rather than minimised (Coelli et al., 2005). Therefore, rather than identifying the closest efficient point on the frontier, it identifies the furthest efficient point on the frontier. Secondly, this model is not unit invariant (Coelli et al., 2005). This means that if the measurement of, for example, one of the inputs were to change from days to hours, this could result in a different efficient frontier being obtained, and as a result, different slacks and weights.

A multi-stage model proposed by Coelli (1998) can be used to avoid the above mentioned problems. However, given that it is a mutli-stage model rather than a two-stage model, it is computationally intensive, as indicated by Coelli (1998).

Koopmans (1951) was one of the first authors to suggest a strict definition of technical efficiency which is equivalent to stating that, in order to be completely efficient, a DMU should have an efficiency score of one and all slacks should be zero. A DMU that meets these criteria is said to be **DEA Efficient** (Sherman and Zhu, 2006). If the DMU only meets the Farrell (1957) definition of efficiency in that its technical efficiency score is one but the slacks are non-zero, the DMU is said to be **weakly DEA efficient** (Sherman and Zhu, 2006).

4.6.4 Incorporating Returns to Scale

The models discussed in Sections 4.6.2 and 4.6.3 all assume constant returns to scale. In many applications it is inappropriate to imply CRS, as different portions of the frontier may exhibit variable returns to scale. Allowing for variable returns to scale in a DEA model, allows scale inefficiencies to be identified (Sherman and Zhu, 2006).

Returns to scale can be varied by using an additional constraint in the linear programming problems 4.2 and 4.3. Varying the constraint will determine the type of returns to scale (Sherman and Zhu, 2006). For example, to impose VRS on the model, the constraint $\sum_{j=1}^{n} \lambda_j = 1$ is added to the list of constraints. Similarly, to impose non-increasing returns to scale (NIRS), the constraint $\sum_{j=1}^{n} \lambda_j \leq 1$ is added. Conversely, to impose non-decreasing returns to scale (NDRS), the constraint $\sum_{j=1}^{n} \lambda_j \geq 1$ is added.

4.6.5 Summary

Tables 4.1 and 4.2 summarise the multiplier and envelopment models discussed above.

Table 4.1: Summary of the Multiplier Model.
Source: Sherman and Zhu (2006).

Frontier Type	Input-Orientated
CRS	$max \ \sum_{r=1}^{s} \mu_r y_{r0} + \mu$ subject to: $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} \nu_i x_{ij} + \mu \leq 0$ $\sum_{i=1}^{m} \nu_i x_{i0} = 0$ $\mu_r, \nu_i \leq 0 \ (\varepsilon)$ where $\mu = 0$
VRS	μ is free
NIRS	$\mu \leq 0$
NDRS	$\mu \geq 0$

Table 4.2: Summary of Envelopment Models.
Source: Sherman and Zhu (2006).

Frontier Type	Input-Orientated
CRS	$min\ \theta - \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s}\right)$ subject to: $\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{i0}$ for $i = 1, 2, ..., m$ $\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{r0}$ for $r = 1, 2, ..., s$ $\lambda_j \geq 0 \ \ j = 1, 2, ..., n$
VRS	$\sum_{j=1}^{n} \lambda_j = 1$
NIRS	$\sum_{j=1}^{n} \lambda_j \leq 1$
NDRS	$\sum_{j=1}^{n} \lambda_j \geq 1$
Efficiency Target	$\hat{x}_{i0} = \theta^* x_{i0} - s_i^{-^*} \ i = 1, 2, ..., m$ $\hat{y}_{r0} = y_{r0} + s_r^{+^*} \ r = 1, 2, ..., s$

4.7 Changes in Efficiency Over Time

In order to analyse panel data, DEA can be used to derive Malmquist Total Factor Productivity (TFP) indices. The use of Malmquist indices to investigate changes in efficiency over time was first investigated by Caves et al. (1982), who used both input and output distance functions to construct two indices. This approach was developed further and generalised for the hospital environment by Fare, Grosskopf, Lindgren and Roos (1989).

The TFP index can be explicitly decomposed into the product of a technical efficiency change index and a technological change index (Coelli et al., 2005). As discussed in Section 3.2.1, technical efficiency can be thought of as the distance that the unit lies from the efficiency frontier, whilst technological change relates to how the efficiency frontier shifts over time (Färe et al., 1994). Furthermore, Färe et al. (1994) suggest that technical efficiency can be decomposed into a “pure” technical efficiency component and a scale efficiency component. The scale efficiency component captures changes between technology under CRS and VRS, whilst the ‘pure’ technical efficiency change denotes the actual change in efficiency over time, without the effects of changes in size of the units. The Malmquist TFP indices are particularly useful in analysing the different components of efficiency over time.

4.8 Using Ratio Analysis to Enhance DEA Results

Ratio analysis is a valuable tool for gaining insights into an organisation, especially when used in conjunction with other methods such as DEA. Furthermore, ratios are particularly useful in scenarios in which no efficiency or production standards exist, as these ratios can be used to estimate the level of production and operating performance (Sherman and Zhu, 2006).

Ratio analysis is typically used to measure the financial performance of a company; however, in the context of efficiency measurement, ratio analysis focuses on the ratio of single inputs to single outputs. The greatest advantages of ratio analysis are the simplicity of calculation and ease of interpretation. Another major benefit of ratio analysis is its flexibility. A variety of outputs and inputs can be selected in order to evaluate different features of the production process (Sherman and Zhu, 2006).

Whilst the majority of hospital efficiency studies use either DEA or SFA, ratio analysis is a useful check on the results obtained from these models. Ratio analysis is a sound method of identifying abnormal relationships between variables; for example very high costs per patient day and low occupancy rates (Sherman, 1984). However, this suggests that a level of judgement is required as to what constitutes an abnormal relationship. When evaluating the efficiency of a group of units, a judgement will need to be made as to what value of the ratio represents an efficient unit. Such judgements are not necessary when using DEA to measure efficiency, as actual efficiency scores are produced. Typically, the distinction between efficient and inefficient units, when using ratio analysis, is arbitrary and, as a result, it is difficult to justify a meaningful cut-off value (Sherman and Zhu, 2006). It is therefore useful to use ratio analysis in conjunction with DEA.

4.9 DEA as a Management Tool

As discussed in Section 4.2, DEA compares different units by analysing the different relative mixes of inputs and outputs, in order to identify the best practice units which use the smallest amount of inputs in order to produce a given level of output, under an assumption of input-orientation (Sherman and Zhu, 2006). An efficiency score is allocated to each unit when it is compared relative to the best practice units; in this way, DEA is able to address productivity by looking at the efficiency of each unit.

By explicitly considering the relative mix of inputs and outputs, DEA allows managers to identify whether a greater number of inputs are being used than is strictly necessary and, given the prices of the inputs, whether the correct mix of inputs is being used (Sherman and Zhu, 2006). In the hospital environment, where hospitals often treat a different mix of cases, DEA allows for adjustments with respect to this case mix to be made, such that the impact of case mix adjustment on productivity can be assessed. Using a case mix adjustment, firms experiencing different mixes in cases can be compared. As a result, DEA can be used to calculate the cost savings that would be achieved, if each of the inefficient units was made as efficient as possible, given the differences in case mix. This is therefore useful for analysing and improving overall profitability.

By using DEA to identify the best practice units, managerial expertise and techniques utilised in the best practice units can be identified and can be transferred to inefficient units in order to improve their relative efficiency (Sherman and Zhu, 2006). This can result in increased efficiency of units, lower operating costs and thus increased profitability and an increase in the quality of outputs produced.

DEA is a constructive tool in industries where there are no industry efficiency standards. For example, it is difficult to develop efficiency standards in the hospital environment as there are a large number of hospitals operating in different locations using a wide range of resources to perform many procedures. Furthermore, the heterogeneity across cases and treatment procedures is too great to develop standards and it requires professional judgment to determine efficiency, which is subjective (Sherman and Zhu, 2006). DEA can be used to identify the best practice units, which can act as the starting point for building standards that reflect the most efficient standards (Sherman and Zhu, 2006). If objective efficiency standards do exist, DEA can be used to evaluate units relative to the existing standard (Sherman and Zhu, 2006).

In summary, DEA is a useful management tool for assessing productivity and profitability. It facilitates the identification and transference of managerial expertise, the establishment of efficiency standards and best practices, pricing and the evaluation of competitive pressures.

4.10 Using DEA to Measure Hospital Efficiency in South Africa

To date, there have only been three DEA studies analysing the efficiency of hospitals in South Africa; all studies looked exclusively at the public hospital sector. Zere, McIntyre and Addison (2001) analysed hospitals in the Northern, Eastern and Western Cape, Kibambe and Koch (2007) focused exclusively on Gauteng and Kirigia, Lambo and Sambo (2000) concentrated on hospitals in Kwa-Zulu Natal. This section comprises a brief overview of these three studies.

Similar inputs were used in all studies. The two inputs used in the Zere, McIntyre and Addison (2001) study were recurrent expenditure, which was a proxy for labour and supplies, and beds, which was a proxy for capital investment. Due to differences in data availability, Kibambe and Koch (2007) were able to use the number of physicians, comprising both doctors and specialists, as well as the number of nurses, rather than using a proxy for labour as in the Zere et al. (2001) study. The number of active beds was also used as a proxy for capital investment in the Kibambe and Koch (2007) study. Outpatient visits and inpatient days were used as outputs in both studies. Kibambe and Koch (2007) use total admissions and total number of surgeries as two other outputs in their analysis. Kirigia et al. (2000) used a larger number of inputs than either Zere et al. (2001) or Kibambe and Koch (2007), including technicians, paramedics, administrative staff, general staff, labour provisioning staff and other staff. None of the studies use a case mix adjustment to account for the different mix of cases seen by each hospital due to unavailability of clinical data.

The sample size across the studies differed significantly. Zere et al. (2001) analysed 86 public hospitals in the Northern, Eastern and Western Cape. The hospitals were classified into three groups consisting of 55, 19 and 12 hospitals, depending on the size of the hospital and the scope of activity. In comparison to the total number of hospitals analysed in Zere et al. (2001), the sample used by Kibambe and Koch (2007) is very small. Kibambe and Koch (2007) started with 29 public hospitals, of which three were excluded as they were specialist hospitals, leaving a sample of 26 hospitals. Of these 26 hospitals, only 14 were able to provide data. Not all 14 hospitals were included because of incomplete information. The available monthly data was restructured into 42 different observations.

Both Zere et al. (2001) and Kirigia et al. (2000) used an input-oriented DEA model to measure technical efficiency. Zere et al. (2001) argued that a hospital's outputs are largely determined by public demand and that hospitals have greater control over inputs in comparison to outputs. Kibambe and Koch (2007) do not specify which orientation was used. All three studies ran models under both constant and variable returns to scale.

Zere et al. (2001) found that only 13% of all hospitals analysed were operating fully efficiently in comparison to their peers. Under constant returns to scale, the minimum technical efficiency scores across the three groups of hospitals ranged from 28.3% to 51.8%. The minimum technical efficiency scores were slighly higher under variable returns to scale, varying across the three groups from 44.2% to 67.1%. Mean technical efficiency scores ranged from 68.1% to 82.8%. Kibambe and Koch (2007) ran a number of different models using different combinations of inputs. Across all the multiple output models, average technical efficiency scores ranged from 70.3% to 98.9%. Kirigia et al. (2000) found that 60% of hospitals were operating fully efficiently, with the other 40% experiencing some level of inefficiency. The outcomes of these three studies are not directly comparable because of differences in the periods analysed, the number of inputs and outputs used, as well as differences in actual inputs and outputs. A major limitation of all three studies was the poor quality of the data available.

Based on the outcomes of their research, Zere et al. (2001) conclude that, given South Africa's budget constraints, improving hospital efficiency by decreasing resource usage is preferable to the government spending large quantities of taxpayer money on an inefficient system. Decreasing the number of beds was recommended for many of the large hospitals. This is a recommendation which may not be practical in reality, given the high demand for health care in the public sector. Rather than recommend bed closures, Kibambe and Koch (2007) suggest that there are too few medical practitioners per active beds because of the high attrition rate of medical professionals. Kibambe and Koch (2007) warn that results should be interpreted with caution, given the small number of hospitals analysed, and emphasised the importance of data collection systems in hospitals. Kirigia et al. (2000) highlight the role that results can play to inform health care decision making, particularly with respect to the allocation and use of health care resources.

Chapter 5

Adjusting for Case Mix

5.1 What is Case Mix?

The term "case mix" refers to the relative proportions of the type or mix of patients treated by a hospital (Fetter, Shin, Freeman, Averill and Thompson, 1980). Individual patients receive different amounts and types of services dependent on their diagnosis and the course of treatment prescribed (Fetter et al., 1980). As a result, hospitals can be viewed as a multi-product organisation where the number of different outputs is as widespread as the number of patients treated (Fetter et al., 1980). A hospital treating a group of patients that require a more sophisticated and expensive set of treatments, such as surgeries, is considered to have a more "severe case mix" or a "heavier case mix" in comparison to a hospital treating a group of patients with minor ailments.

The term "case mix" is also a generic term used to describe statistically developed grouping mechanisms which are used to group patients, in order to assist the planning and management of health care (Heavens, 1999). These are also known as "case mix groupers". "Case mix groupers" can facilitate a number of hospital management activities (Heavens, 1999):

- Budget allocation
- Benchmarking of cases
- Cost management
- Pricing and billing
- Contract funding and performance evaluation
- Quality management
- Establishing community-based burden of disease
- Policy and planning activities

5.2 Diagnosis-Related Groups as a Measure of Case Mix

There are a number of measures of case mix that have been developed and used across the world, the most popular being Diagnosis-Related Groups (DRGs). DRGs are used across a variety of countries, however the specific design and development of the groups vary across regions (Scheller-Kreinsen et al., 2009).

DRGs are a patient classification system which was developed in the 1970s as a management tool to improve the control of hospital costs (Erlandsen, 2008). DRGs have been used in a number of contexts since then, including prospective reimbursement systems in the United States of America and other regions, where hospitals are paid an amount based on the actual mix of cases treated, using prospectively determined amounts based on the expected resource utilisation of each case.

Fetter et al. (1980) note that there are particular demographic, diagnostic and therapeutic attributes which determine the type and level of health care provided. The aim of DRGs is to aggregate similar patients, using these identified attributes, to produce clinically and economically homogeneous groups, such that the cases within each group have similar patterns of resource consumption (Erlandsen, 2008).

Fetter et al. (1980) outline the five requirements of DRGs:

1. DRGs need to be medically interpretable, such that medical practitioners are able to recommend a particular treatment protocol for groups of patients.
2. All the variables used to define the DRGs should be readily available and relevant to resource utilisation.
3. The groups should be mutually exclusive and exhaustive. However, this is balanced against there being a manageable number of groups.
4. Each group should contain patients with similar expected resource usage.
5. Groups should be comparable across different hospital coding systems.

Upon scrutinising these five objectives, it is clear that there is a trade off between medical interpretability and maximising explained variation. Subdividing a group further, to increase the homogeneity and to maximise explained variation of each group, needs to be balanced against the medical interpretation of each group, as well as the total number of groups that is considered to be manageable in the particular scenario. A detailed description of the methodology of constructing DRGs is provided in Fetter et al. (1980).

Fetter et al. (1980) succinctly summarise the usefulness of DRGs in the clinical and financial aspects of health care:

> "groups can provide a framework for the initiation of an ongoing process of comparative analysis of health care with the long run goal of determining both

> the cost and value of any kind of care that might be delivered. With such information, meaningful dialogue among clinicians, administrators, planners and regulators can proceed in rationalising observed differences".

The major drawback of using DRGs to construct a case-mix index is the presence of heterogeneity of resource utilisation within the groups. Initially severity of illness was not taken into account (Linna, 1998), which resulted in significant heterogeneity. For example, consider two women who both have caesarian sections and who both experience complications. The complications for the one may be minor, whilst for the other may be significantly worse. However, both cases are grouped into one category and there is no differentiation between the two cases despite resource consumption being very different. These two cases are likely to use different quantities of resources and are not entirely homogeneous. In more recent years, subcategories relating to levels of complication, prescence of co-morbidities, severity, age and gender have been incorporated. Whilst the heterogeneity is not as significant as it used to be, even within these granular groupings, heterogeneity may still be present.

5.3 Rationale for Adjusting for Case Mix

Few of the existing studies examining hospital efficiency adjust the outputs used to reflect differences in case mix across hospitals. Effectively, this assumes that the case mix across hospitals is uniform, in that each hospital treats an identical mix of cases. This is an oversimplification of reality and is likely to influence efficiency scores. Consider two hospitals, one which has no operating theatres and only treats minor cases and another which has multiple theatres and specialises in heart conditions. The types of cases being treated in each hospital will be very different and will require different types and amounts of resources. It is thus unreasonable to assume that all hospitals treat the same mix of patients.

In order to make efficiency scores comparable across different types of hospitals, which operate in different locations and treat a different mix of patients, an adjustment needs to be made to account for the differences in case mix. If no adjustment was made, a hospital treating a more severe mix of patients than the average hospital will be using significantly more resources to treat the same number of people and, as a result, will appear to be relatively inefficient. However, the additional resource utilisation should be attributed to differences in case mix, rather than inefficiency. A similar argument can be made for hospitals that treat a relatively less severe mix of patients. To avoid case mix confounding efficiency scores, the data needs to be adjusted such that hospitals are effectively treating a standard mix of cases.

Currently, there is no standard method to incorporate differences in case mix into a DEA model. Gustafson and Holloway (1975) suggested that, at the time, most methods that had been proposed had not been widely adopted, because medical professionals did not understand how the methods worked or how to apply them. Arguably, this is still the case. As a result, there are a variety of techniques available but there is no single commonly accepted method. Three main methods are discussed below, namely disaggregation of admissions, additional output variables, and an adjustment to outputs using case-mix adjustment factors.

5.4 Disaggregation of Admissions

Sherman (1984) and Grosskopft and Valdmanis (1993) suggest disaggregation of admissions as a method for accounting for variation in resource utilisation, as a result of differences in case mix.

Since DEA is able to accommodate multiple outputs, Sherman (1984) suggests that grouping cases using relative weights to take account of case mix, which is discussed in Section 5.6, is unnecessary. Instead, case mix can be considered explicitly by using disaggregated outputs; for example, by including the number of cases in each DRG as an individual output. However, Sherman (1984) does not consider the effect of dramatically increasing the number of output variables and the impact this has on efficiency scores, as described in Section 4.3. Whilst disaggregating admissions into broad categories is unlikely to significantly impact efficiency scores, using each DRG as a separate output is not feasible, given the very large number of groups.

As a result of a lack of detailed information, the DEA model proposed by Sherman (1984) divides patient days into patient days for patients younger than 65 and patient days for patients who are older than 65. Ideally, a more detailed set of outputs would be used to capture case mix, however the data set prevented it (Sherman, 1984). Although age is a key factor in determining relative resource utilisation, age alone offers an incomplete measure of case mix (Sherman, 1984). As a result, efficiency scores may be biased because other dimensions of case mix have not been considered or captured in the disaggregation process. This is an important consideration when deciding upon how best to disaggregate outputs to reflect case mix.

Grosskopft and Valdmanis (1993) go a step further than Sherman (1984) and compare the disaggregation of admissions to admissions adjusted for case mix, using a case mix adjustment factor. The construction and use of a case-mix adjustment factor is described in detail in Section 5.6. In the disaggregated output model, hospital admissions were disaggregated into maternity, emergency, surgical and outpatient cases. Whilst this allows for some case mix variation to be captured, significant variation may still exist within each of these categories, most notably the surgical category (Grosskopft and Valdmanis, 1993). For example, consider an appendectomy and open heart surgery. These two procedures will require different lengths of time under anesthesia, different lengths of stay, different acuity of care and will have a range of different complications. The case-mix adjustment factor model used an identical set of inputs and outputs, except that the disaggregated outputs were multiplied by the case mix adjustment factor. Upon comparing the two models, there appears to be no significant difference between the efficiency scores produced by the two different models (Grosskopft and Valdmanis, 1993). This is a significant finding, as it suggests that the case-mix adjustment factor does not capture any differences in case mix, over and above what is captured through a broad disaggregation of admissions. This is a particularly useful result in cases where there is little detailed data available and the construction of a case mix adjustment factor is not possible. It is important to note that this finding is likely to vary among data sets.

5.5 Additional Outputs

An alternative to disaggregation is including an additional output variable which provides information about relative resource consumption, as a result of differences in case mix.

Grosskopft and Valdmanis (1993) suggest that the case-mix adjustment factors can be included as an additional output. However, when analysing results of the technical efficiency, these results were not significantly different from the results that contain no adjustment for case mix (Grosskopft and Valdmanis, 1993). This indicates that using the adjustment factors as an additional output does not necessarily capture the differences in case mix. The construction of a case-mix adjustment factor is detailed in Section 5.6.

Zuckerman, Hadley and Iezzoni (1994) propose the inclusion of an output which accounts for the number of high technological services available in each hospital. High technological services include cardiac catheterisation laboratories, open heart surgery facilities, extracorporeal shock-wave lithotripters, megavoltage radiation therapy, nuclear magnetic resonance imaging, organ/tissue transplant centres, and certified trauma centres (Zuckerman et al., 1994). The rationale for including these high technology services is that the more high technological services a hospital has, the more likely the hospital is to attract sicker patients who will require more resources (Zuckerman et al., 1994). The drawback of this method is that detailed operational information is required for each hospital analysed. That said, operational information is likely to be easier to get in comparison to detailed clinical information relating to each patient treated.'

5.6 Adjusting Outputs Using a Case-Mix Adjustment Factor

The most popular method in international literature to adjust outputs to reflect differences in case mix, is to multiply the outputs by a case-mix adjustment factor. This method is used in Clement et al. (2008), Mutter et al. (2008), Linna (1998), Zuckerman et al. (1994), Grosskopft and Valdmanis (1993), Fetter (1991) and Fetter et al. (1980). The aim of a case-mix adjustment factor is to capture differences in expected resource utilisation across types and severity of illnesses, such that, when comparing efficiency scores, the extent to which differences can be attributed to patient characteristics or to treatment practices can be separated out. Although the overall method is the same across these studies, the techniques used to construct the case-mix adjustment factors are varied.

The majority of the studies construct factors based on the distribution of cases across DRGs (Linna, 1998; Zuckerman et al., 1994; Fetter, 1991; Fetter et al., 1980). Few studies make use of alternative methods. The expected resource consumption for a hospital for each DRG, is compared against the average resource consumption across all hospitals for each DRG, in order to derive a case-mix factor. Outputs for each hospital are then adjusted by multiplying each output by the respective case-mix adjustment factor.

Fetter et al. (1980) use patient length of stay (LOS) as a proxy for resource consumption and analyse the difference between the average LOS for a hospital and the average LOS across the hospital group, before deriving a set of case-mix adjustment factors. The difference between the average LOS in hospital i (a_i) and the average LOS for the industry (A) can be broken into three components, namely the difference due to hospital specific factors, differences that are attributable to differences in case mix, and an interaction component.

Consider:

$$a_i = \sum_j p_{ij} a_{ij} \text{ for } j = 1, 2, 3, \ldots \tag{5.1}$$

$$A = \sum_j P_j A_j \text{ for } j = 1, 2, 3, \ldots \tag{5.2}$$

where

a_{ij} = the average LOS in the j^{th} DRG for hospital i,

p_{ij} = the proportion of hospital i's cases in the j^{th} DRG,

P_j = the proportion of all hospital cases in DRG j,

A_j = the average LOS across all hospitals for the j^{th} DRG,

a_i = the actual average LOS in hospital i across all DRGs, and

A = the actual average LOS for all hospitals across all DRGs.

$$
\begin{aligned}
\underbrace{a_i - A}_{\text{Average LOS Difference}} &= \underbrace{\left(\sum_j P_j a_{ij} - \sum_j P_j A_j\right)}_{\text{Hospital Related Factors}} + \underbrace{\left(\sum_j A_j p_{ij} - \sum_j A_j P_j\right)}_{\text{Case Mix}} \\
&\quad + \underbrace{\sum_j \left(a_{ij} - A_j\right)\left(p_{ij} - P_j\right)}_{\text{Interaction Component}} \text{ for } j = 1, 2, 3, \ldots
\end{aligned}
\tag{5.3}
$$

The hospital-related factors component is the difference in the average LOS that can be attributed to the difference between the hospital specific LOS and the overall average LOS, using a constant industry mix of cases. This difference may be a result of different treatment protocols, different equipment available for use and differences in the proficiency of staff, amongst other factors. This is the component that can be attributed to differences in efficiency.

The case mix component is the difference in the average LOS that can be attributed to differences in case mix across hospitals. Using the average LOS for each DRG across the whole industry, this component measures the difference between one hospital's mix of cases and the industry average. Hospitals have little control over this factor and differences should be adjusted for in an efficiency analysis.

The interaction component is the difference in the average LOS that cannot be attributed to, either differences in hospital specific factors, or to differences in case mix. Fetter et al. (1980) warn that if there is a large positive interaction component, the standardisation may be misleading and this method should be used cautiously.

The Case-Mix Adjustment Factor (CMAF) for hospital i can be calculated as the case mix adjusted LOS for hospital i divided by the average LOS for all hospitals (Fetter et al., 1980):

$$
CMAF_i = \frac{\sum_j A_j p_{ij}}{\sum_j A_j P_j} \text{ for } j = 1, 2, 3, \ldots \tag{5.4}
$$

A large CMAF value indicates that hospital i has a relatively more resource intensive mix of cases, in comparison to the average of all hospitals. Whilst the initial analysis of the average LOS difference is not directly needed to calculate the case-mix adjustment factor, it does provide hospital specific information regarding differences due to case mix and differences due to hospital related factors. Furthermore, it is necessary to analyse the interaction component to identify hospitals for which standardisation may not be appropriate.

Rosko and Chilingerian (1999) suggest a similar method of calculating a case-mix adjustment factor. However, rather than using LOS as a proxy for resource utilisation, total costs are used. The CMAF for hospital i can be calculated as follows:

$$CMAF_i = \frac{\sum_{j=1}^{n} w_j p_{ij}}{\frac{1}{n}\sum_j \sum_i w_j p_{ij}} \tag{5.5}$$

where,
w_j = the relative costliness of the j^{th} DRG,
p_{ij} = the proportion of hospital i's cases in the j^{th} DRG, and
n = the number of hospitals.

An identical formula was presented in Rosko and Carpenter (1994). The interpretation of the value of the Rosko and Chilingerian (1999) CMAF is identical to the Fetter et al. (1980) CMAF. The major difference between these two proposed case-mix adjustment factors is the measure of resource utilisation. Fetter et al. (1980) suggest the use of LOS as a proxy for resource utilisation, whilst Rosko and Chilingerian (1999) use the actual costs incurred. Both measures have been used in international hospital efficiency literature to compare the performance of hospitals, with respect to resource utilisation (Fetter et al., 1980). The advantage of both these measures is that the data are often readily available and they are easy to interpret (Fetter et al., 1980; Rosko and Carpenter, 1994). However, LOS could be considered more appropriate than actual costs incurred, when performing an efficiency analysis. Costs may vary between hospitals and across geographic locations, particularly nursing costs (O'Neill et al., 2008). If actual costs are used, these differences would be classified as differences in efficiency, rather than differences that cannot be controlled by the hospital. The disadvantage of using LOS is that it assumes that a day in hospital will use the same level of resources, irrespective of the diagnosis of the patient. Consider the resource utilisation for a day in an Intensive Care Unit (ICU) in comparison to a day in the general ward, the resource utilisation in ICU will be significantly higher. Given the advantages and disadvantages of these two proxies of resource utilisation, use of either of these proxies can be argued. Unlike Fetter et al. (1980), Rosko and Chilingerian (1999) do not warn against standardisation for certain hospitals.

Rosko and Chilingerian (1999) argue that intra-DRG variations in severity of illness and resource utilisation should be accounted for when determining hospital cost functions and measuring hospital efficiency. If these variations are not taken into account and significant variations in resource utilisation exist within groups, these variations will be classified as inefficiencies, while in reality it can be attributed to differences in severity of illness. It has been argued that large variations in severity of illness are not problematic, as the distribution of resource utilisation of a group is highly peaked with thin long tails (Horn, Sharkey, Chambers and Horn, 1985). Horn et al. (1985) argue that if the cases in the tails of the distribution, in other words the very severe and less severe cases, are randomly distributed across hospitals, then the variation is not likely to be a problem. However, there is a possibility that severe and less severe cases may not be randomly distributed across hospitals because of differences in hospitals' expertise and ability to handle complex cases. Linna (1998) suggests that the effect of the variation in intra-DRG severity of illness is difficult to ascertain when DRGs are used as a method of aggregating cases into a single output.

5.7 Adjusting Existing Efficiency Studies to Allow for Differences in Case Mix

Many of the existing DEA hospital efficiency studies do not account for case mix because, at the time of writing, sufficiently detailed data was unavailable. In order to better understand and compare results, it would be useful to be able to account for case mix after the fact. Little research has been carried out in this field.

A Nordic study of the impact of ownership reform on hospital efficiency uses a DEA model to determine efficiency scores and then uses a regression model to determine the impact of factors that are not inputs and outputs on the efficiency scores, such as the impact of case mix (Kittelsen, Magnussen, Anthun, Häkkinen, Linna, Medin, Olsen and Rhenberg, 2008). Two different variables are used as regressors, namely a case-mix index and a length of stay deviation, which is the difference between the actual length of stay and the expected length of stay. In an email on 10 November 2010, one of the authors of this paper, Clas Rehnberg, suggested using a case-mix index to rank hospitals. This rank can then be used as a regressor. This is an alternative potential method for determining the impact of case mix after having evaluated efficiency scores.

A similar method is used by Zere et al. (2001) to investigate the impact of the size of a hospital on efficiency scores after the fact. However, Zere et al. (2001) use a Tobit model, also known as a censored regression model, rather than a standard regression model.

Chapter 6

Data and Methodology

6.1 Data

6.1.1 Data in the Hospital Industry in South Africa

The aim of this dissertation is to measure the efficiency of hospitals operating in the private hospital industry in South Africa and to determine the impact of controlling for case mix on efficiency scores. One of the major reasons for focusing solely on the private hospital industry, is the poor availability and quality of data in the public sector, as highlighted by Kibambe and Koch (2007). Furthermore, because of the concentration of ownership in the private hospital sector, problems of consistency across data sets are reduced and data relating to multiple hospitals can be obtained from a single source.

Although this dissertation is focused solely on the private hospital industry, given an appropriate data set relating to the public sector, the methodology discussed below could be applied to this data set to evaluate the relative efficiency of the public hospital industry.

6.1.2 Data Requirements

The data required for this investigation comprises the data needed to measure the chosen inputs and outputs, for each of the hospitals within this specific hospital group, for each year of the analysis. Data are also required to construct the case-mix adjustment factor.

6.1.3 Source of Data

Data were obtained from one of the three major private hospital groups operating in the South African private hospital sector. The group provided clinical, human resource and operational data for the 53 hospitals they owned, for years 2007 to 2011 inclusive.

Sourcing data from a single private hospital group has the advantage of the data being consistently recorded across all the hospitals in an identical manner. All fields are available for all the hospitals. Furthermore, there is no variation in coding systems and definitions across the hospitals. This removed many of the difficulties of working with multiple data sets from different sources, particularly when calculating a case-mix adjustment factor which requires consistent grouping of cases into DRGs or other clinical groups.

The major limitation of this data source is that it is not necessarily representative of the South African private hospital industry as a whole. However, the data set is large and covers approximately a third of the industry. Therefore, the results of this analysis are expected to provide an initial insight into the efficiency of the private hospital sector in South Africa. Because of significant structural differences between the private and public sector hospital environment, these results are not applicable to the public hospital sector in South Africa, but may be used as a basis of comparison at a later stage.

6.1.4 Characteristics of the Data

Detailed clinical, human resource and operational data were provided for each of the hospitals owned by the private hospital group for each of the years from 2007 to 2011.

The clinical data was provided at case level, for each hospital, for each of the years. The case-level data consisted of the following data fields:

- admission date and time
- the discharge date and time
- gender
- date of birth
- the medical scheme name and code
- an account category which broadly describes the case type, for example a general medical admit or a surgical day case
- a number of case mix classification codes, such as Diagnostic-Related Groups (DRGs), Basic-Diagnostic-Related Groups (BDRGs) and Major Diagnostic Categories (MDCs)
- the total billed amount
- the total pharmacy amount
- calendar days spent in hospital
- number of bed days sold
- theatre minutes for a major procedure, for a minor procedure and for the catheterisation laboratory (cath lab)
- the expected mortality rate of the patient
- the actual mortality rate of the patient

- a number of case mix adjusted indices for the total billed amounts and the pharmacy amounts

Data relating to the calculation of the case-mix adjustment factor was only available from 2008 onwards. More specifically, the classification codes, as well as some of the case mix adjusted indices for the total billed amounts and the pharmacy amounts, were not available for 2007. The same grouper was used across the full four year period.

The human resource data were provided on a monthly basis, for all the hospitals, across all the years. The data consisted of a job category, the number of employees employed each month under this job category, the average salary paid to employees across each job category, as well as the sum of salaries paid across a single job category. It is important to note that hospitals are not allowed to employ doctors (Matsebula and Willie, 2007), therefore the human resource data provided pertains to nursing staff and other support staff. No data were available pertaining to the number of doctors working in a particular hospital or the number of hours worked.

Nursing agency staffing data were also supplied. Certain hospitals appeared to make extensive use of agency staff. This was discussed with the hospital group and it emerged that these hospitals make use of agency staff for two distinct reasons. Firstly, some hospitals use agency staff during very busy periods when the existing staffing level is inadequate, rather than increasing the number of employees, in order to minimise fixed costs. Secondly, in the more rural areas, there are often a number of hospitals but very few nurses. As a result, nurses often work on an agency basis, such that they are able to work across a number of different hospitals. The agency data provided included the hospital name, the month, a job category description, a description of the shift worked (such as standby, call out or shift worked), the number of hours or units worked, the rate per hour or unit, and the total client value.

The operational data were provided for each hospital for a number of months each year. The data consisted of the number of licensed beds, the number of beds in operation, the number of available beds, the number of licensed theatres, as well as the number of theatres in operation.

As complete clinical data were not available for some hospitals for the full year in 2007, the year 2007 was necessarily excluded from the analysis. The final period analysed consisted of the four years from 2008 to 2011.

6.1.5 Data Cleaning

Various adjustments were made to the data set during the data-cleaning process. Adjustments were made on two different levels: the hospital level and the case level.

At the hospital level, a number of hospitals were excluded from the analysis, leaving a total of 41 hospitals in the final sample. All hospitals operating outside South Africa were excluded as the focus of this research is the South African market. In addition, five hospitals had either incomplete clinical data sets or missing human resource data, as they had been incorporated into the hospital group during the study period. These five hospitals were removed from the analysis to avoid confounding comparisons of efficiency by including hospitals with incomplete information. There was no indication that these hospitals are atypical when compared to the other hospitals. Of the remaining hospitals, two hospitals are day hospitals and another a specialist hospital. These three hospitals are atypical when compared to the other hospitals in terms of resource utilisation and the types and number of cases treated. To prevent a distortion of the results, these three hospitals were excluded from the final data set. Two of the remaining hospitals were small and in close proximity to one another and operate under the same management team. For the purposes of this analysis, these two hospitals were merged. The total sample size after data cleaning was 41 hospitals. All hospitals were de-identified and were ordered from the smallest to largest, by number of operational beds in 2011, and then renamed from 1 to 41.

At a case level, only those cases which resulted in a hospital admission were included in the analysis. Examples of the cases excluded were pharmacy-only cases, theatre-only cases and cases involving only the use of catheterisation laboratories. These cases, known as partial accounts, are atypical when compared to normal hospital admissions, such as maternity cases or medical cases, and often have volatile billed amounts. Inclusion of such cases is likely to distort results, therefore these cases were excluded from the analysis. Cases with either a zero or a negative billed amount, which are usually as a result of reversing incorrect entries into the system, as well as cases used purely for management control purposes, for example fictional cases designed to test the administration system, were also excluded from this analysis.

Spot checks were carried out to ensure the consistency of all the excluded cases. On average, the number of zero and negative billed cases were about 2% of all cases for each hospital. This proportion remained fairly constant across all four years, as well as across each hospital. Partial accounts made up a very small proportion of the number of cases.

Although checks were not carried out on all the cases that were excluded, the spot checks used provided some confidence that the cases which were excluded, were removed appropriately and consistently across years and hospitals. Furthermore, the use of annual data means that the impact of a single case on the annual value is small and random fluctuations are reduced.

Prior to the data cleaning process, many of the cases in the data set had surprisingly small billed amounts, where the total billed amount was very close to zero. After making the above adjustments to the data set, the billed amounts were re-analysed. It was found that the above cleaning procedure had removed the majority of these cases. This provides confidence that, of the cases that remain in the data set, very few have atypically small billed amounts. This provided confirmation that the cleaning process was appropriate. No adjustment was made for cases with very large billed amounts. As such cases can and do occur in the normal hospital environment, it was decided to leave these cases in the data set.

6.1.6 Data Quality

The data set is believed to be of good quality and consistent across hospitals, since they were sourced from a single hospital group with a central data warehouse. As a result, the models should provide a reliable indication of the inefficiencies present in this group of hospitals, and a potential indication of the inefficiencies present in the wider South African private hospital industry.

Data are believed to be consistent across hospitals, since they were sourced from a single hospital group. The coding system used to classify cases is consistent across all hospitals, therefore the coding of cases is consistent. However, it is important to note that there may be some variation in the capturing of clinical codes due to human error. This variation is believed to be negligible.

6.1.7 Data Limitations

In spite of the data being detailed, particularly in comparison to the three existing studies on hospital efficiency in South Africa, namely the studies by Kibambe and Koch (2007), Zere et al. (2001) and Kirigia et al. (2000), this study could be improved by having access to more comprehensive data.

On the input side, the study could benefit from more detailed data relating to capital investment, labour, and goods. Whilst the number of operational beds are likely to be highly correlated to the amount of capital invested, beds do not provide an indication of the *flow* of capital or of fixed-asset turnover. Detailed data relating to the capital flow of each hospital could facilitate the analysis. Furthermore, detailed data relating to medical professionals, other than nurses, would be useful to create a more complete labour input.

On the output side, a measure of quality of care would facilitate the analysis. For the purposes of this analysis, quality of care was assumed to be constant across all hospitals. Given that all hospitals are owned by the same hospital group and are therefore subject to the same management procedures and quality standards, this was assumed to be a reasonable assumption. However, this is not ideal, and detailed quality of care data would benefit the analysis.

The exclusion of five hospitals due to incomplete data may significantly alter the results of the model if included. Since DEA is a relative benchmarking technique, results are dependent on the sample used and there is the possibility that relative efficiencies change if the sample group changes. However, since these five hospitals are a small proportion of the total sample, inclusion of these hospitals at a later date is not expected to significantly alter the results obtained.

As previously noted, the data set relates to hospitals owned by one hospital group. The study could be improved if data could be obtained from the other two major hospital groups, as this would provide a more realistic indication of the inefficiencies present in the private hospital industry as a whole.

6.2 Methodology

6.2.1 Overview of Methodology

As discussed in Section 3.5.4, DEA was chosen as the technique to measure hospital efficiency. There are four major methodological decisions to consider before specifying a DEA model. These decisions include the choice of inputs and outputs, the method of measuring case mix, the orientation of the model, as well as returns to scale. These decisions are discussed in detail below. The software used to complete this analysis is also discussed briefly.

6.2.2 Choice of Inputs and Outputs

As discussed in Section 4.2, a balance needs to be struck between selecting the variables that adequately capture the nature and complexity of the production process and only using a small set of variables, such that the power of the DEA model is retained. In typical studies of hospital efficiency using DEA, there are three major input categories: capital (which includes beds, infrastructure and medical equipment), labour, and goods (Nguyen and Coelli, 2009; Worthington, 2004; Hollingsworth, 2003; Grosskopf and Valdmanis, 1987).

The four inputs examined in the preliminary analysis are:

- the number of operational beds,
- the number of operational theatres,
- the number of nurses adjusted for seniority, and
- an index for pharmacy goods of actual usage over expected usage.

Operational beds and operational theatres are used as a proxy for capital investment. Although neither of these variables are complete and comprehensive measures of capital, they are the best available proxy in the data set. The number of nurses, adjusted by salary for seniority and hence expertise, is an indication of the labour required to produce hospital services. An index of the actual expenditure on pharmacy goods over the expected value adjusted for case mix, is used as a proxy for the goods that are utilised in the production of medical outcomes.

The choice of outputs is more complex. Rather than using final health outcomes, which are difficult to quantify and are not readily available, intermediate health outcomes were used. The following outputs are analysed in the preliminary analysis of the data:

- total theatre minutes, and
- the number of admissions.

These four inputs and two outputs are described in detail below.

Operational Beds

The number of operational beds was used as a proxy for capital investment. Using beds as a proxy for capital investment is consistent with much of the international literature (Nguyen and Coelli, 2009; O'Neill et al., 2008; Worthington, 2004). Parkin and Hollingsworth (1997) note that the rationale for using the number of operational beds in efficiency studies is that the number of beds is an attempt to measure the quantity of capital. Furthermore, the number of beds should be roughly proportionate to the cost of building and maintaining a hospital.

It is important to distinguish between the number of registered beds and the number of operational beds. The number of registered beds is the number of beds that a hospital is licensed to operate, given the staff and resources available. Operational beds are the number of beds that are staffed and available for patients. When analysing the data, it was found that for some hospitals the number of registered beds was larger than the number of operational beds. In order to avoid providing a false impression of a hospital's capacity to admit and treat patients, the number of operational beds was used instead of the number of registered beds.

The number of operational beds for each hospital was provided for a number of calendar months each year. Although changes in the number of beds available do not occur regularly, in some hospitals there were a number of changes in particular years. For this reason, the weighted average number of beds was used, rather than the number of operational beds at the end of the calendar year, to prevent overestimating or underestimating a hospital's ability to accommodate patients. For example, hospital 20 started with 90 operational beds in January 2009, which increased to 126 operational beds in February and to 130 operational beds in October 2009. Using a figure of 130 operational beds for 2009 overestimates hospital 20's ability to accommodate patients over the whole year. A weighted average of 124 operational beds is a more realistic representation. As a result of using a weighted average of the number of beds, it is possible to have fractions of beds.

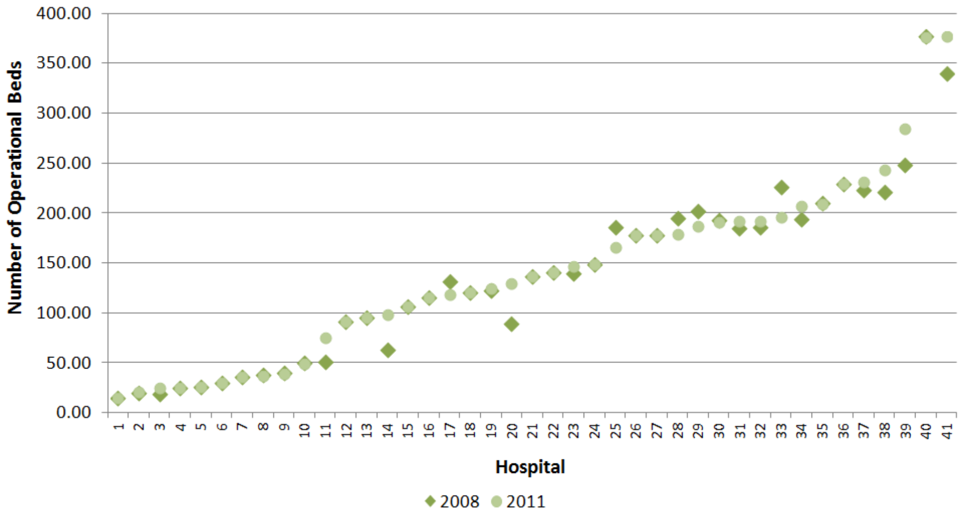

Figure 6.1: Number of operational beds per hospital in 2008 and 2011.

The number of operational beds for each hospital for 2008 and 2011 is illustrated in Figure 6.1. The hospitals are sorted by number of operational beds in 2011. The cutoff of 150 beds was used to divide the group into small hospitals and large hospitals. This cutoff resulted in approximately half the hospitals being classified as large hospitals, and the other half as small hospitals. A further reason for using this cutoff value is that no hospitals changed size classification over the four year analysis period. Hospital 1 to Hospital 24 are classified as small hospitals and Hospital 25 to Hospital 41 are classified as large hospitals for the purpose of this analysis. Of the small hospitals, ten have 50 or fewer beds. There were 15 hospitals that increased the number of operational beds available over the four year period and six hospitals experienced a decrease in the number of operational beds. For the remaining 22 hospitals, the number of operational beds remained constant over the four year period. Twelve large hospitals experienced a change in the number of operational beds, whilst only nine small hospitals experienced a change. The largest absolute increase over the four year period was an increase of 40 operational beds for Hospital 20. Similarly Hospital 33 experienced the largest absolute decrease of 30 operational beds.

An overall occupancy rate was calculated for all the hospitals by dividing the total number of billed days by the total number of bed days available. The occupancy rate for the sample increased from 65.8% in 2008 to 70.1% in 2011. In 2008, the occupancy rate for the private hospital sector was 65.5% (Childs, 2009). This further suggests that the sample used is representative of the private hospital industry.

Individual occupancy rates were calculated for each hospital. The minimum occupancy rate for 2011 is 54% for Hospital 1. Hospital 23 had the maximum occupancy rate for 2011 at 87%. Only 11 hospitals experienced a decrease in occupancy between 2008 and 2011. Eighteen hospitals experienced a more than 10% increase in occupancy rates over the four year period. Occupancy rates for 2008 and 2011 are illustrated in Figure 6.2. Occupancy rates do not appear to vary materially with the number of beds.

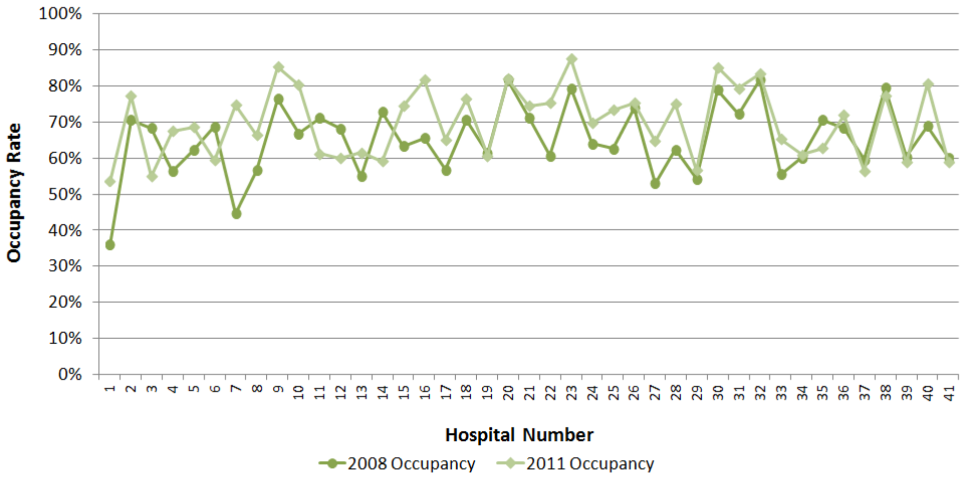

Figure 6.2: Occupancy rates for 2008 and 2011.

Three (Hospital 14, 11 & 3) of the four hospitals that experienced an over 30% change in the number of operational beds, experienced the largest decrease in occupancy rates over the four year period. The likely reason for this decrease in occupancy rates is that, despite increasing the number of beds available, the demand for hospital care has not increased by the same proportion. Of these four hospitals, only Hospital 20 experienced a minor increase in occupancy rates. There were 22 hospitals that did not experience a change in the number of beds. Of these 22 hospitals, two hospitals experienced a decrease in occupancy rates (Hospital 6 & 35). These two hospitals are of particular concern, as the number of bed days sold has decreased substantially over the four year period. Another potential reason for changes in occupancy rates is the establishment of new hospitals within the local market, therefore increasing competition and the supply of beds, resulting in a decrease in bed occupancy within the local market.

Hospitals with an occupancy rate of 80% are considered to be operating at full capacity (Childs, 2009). An occupancy rate of 100% is neither practical nor possible due to the strain it places on support services, the increased risk of medical errors as well as the increased risk of hospital-related infections (Childs, 2009). In 2008 only two hospitals had an occupancy rate of over 80%. This increased to eight hospitals in 2011. This suggests a growing demand for private hospital services and could be a result of the increase in the population covered by medical schemes.

A useful extension to measuring operational beds would have been to look at the split of beds by ward and adjust the total number of beds in the same way as which the total number of nurses was calculated. However, at the time of writing no data was available regarding of the split of beds between different wards.

Operational Theatres

The number of operational theatres was investigated as a second proxy for capital investment. As with beds, the number of operational theatres was used rather than the number of licensed theatres. The weighted average number of operational theatres was calculated for 2008, 2010 and 2011, for the same reasons as the weighted number of operational beds was calculated. No theatre data were available for 2009. The average of the number of theatres for 2008 and 2010 was used as an approximation for the number of theatres in 2009.

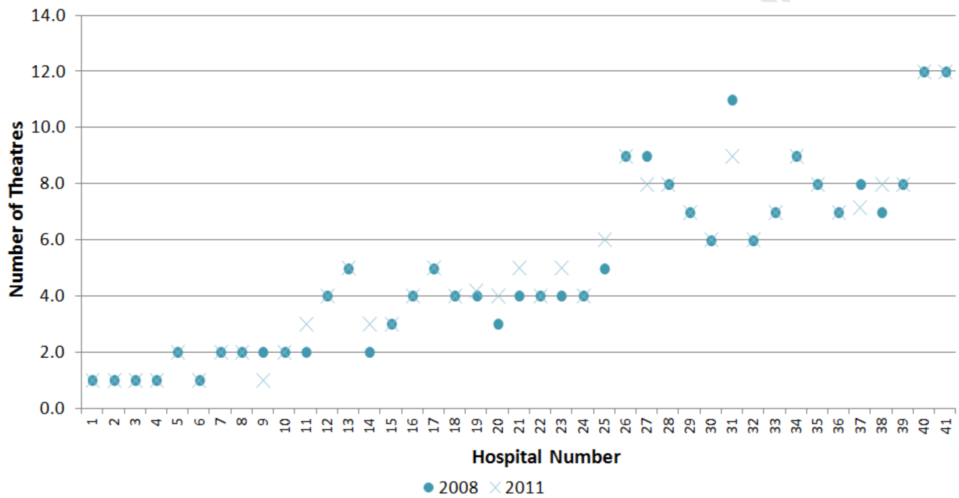

Figure 6.3: Number of operational theatres per hospital in 2008 and 2011.

Figure 6.3 illustrates the number of operational theatres for each hospital, for both 2008 and 2011. The hospitals are sorted using the number of operational beds for 2011. In general, the number of operational theatres increases with the size of the hospital. Four of the 41 hospitals experienced a decrease in the number of operational theatres, whilst eight of the 41 hospitals experienced an increase in the number of operational theatres. The largest decrease was two theatres over the four year period for Hospital 33. The two largest hospitals have twelve operating theatres each over the four year period. The ten smallest hospitals have no more than two operational theatres from 2008 to 2011.

The use of theatres was measured through a theatre occupancy rate. This was calculated by assuming a theatre will be used for eight hours a day, seven days a week as suggested by Costa, Buys and Pyle (2009). This takes into account the time that theatres are not available for use during post-operative cleaning and the preparation time for a surgery (Costa et al., 2009). These calculations indicate that 24 hospitals in 2008 were used for more than eight hours a day for surgical procedures and only 20 hospitals in 2011. Over the four year period, 20 hospitals experienced an increase in utilisation.

The correlation between operational beds and operational theatres was checked. As discussed in Section 4.2, little information is gained from including a highly correlated variable into a model, and it may bias efficiency scores.

It is clear in Figure 6.4 that there is a linear relationship between operational beds and operational theatres. This relationship is not completely unexpected as the number of theatres should increase with the size of the hospital, as measured by operational beds. Pearson's correlation coefficient was calculated and has a value of 0.945.

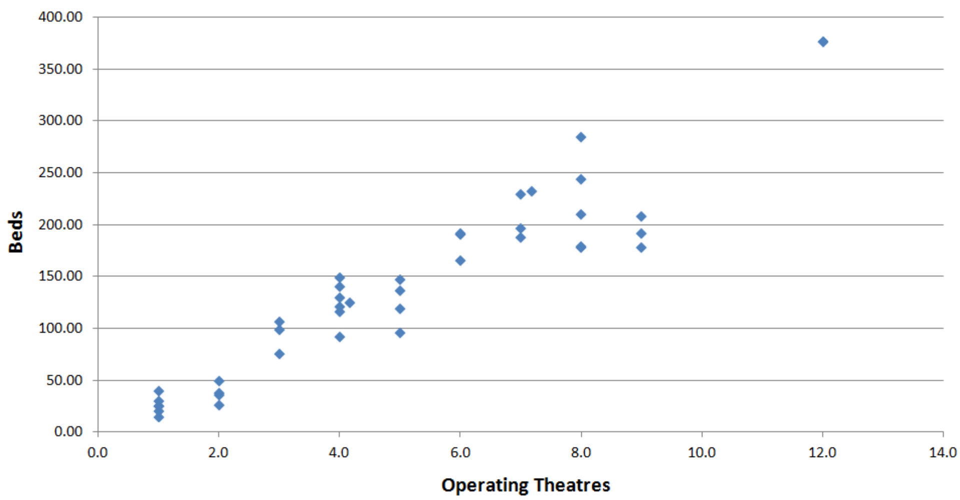

Figure 6.4: Correlation between operational beds and operational theatres in 2011.

Given the relationship displayed in Figure 6.4, as well as Pearson's correlation coefficient, it may be suggested that operational theatres do not offer a great deal more information than operational beds. For this reason, the number of operational theatres is not included in the final model used in this analysis.

Adjusted Number of Nurses

The number of nurses, adjusted by salary to reflect seniority and expertise, was used as a proxy for labour in the production process. As South African hospitals are prohibited from employing medical practitioners with the exception of nursing staff (Matsebula and Willie, 2007), the human resource data set obtained provided no information relating to doctors or specialists working within any of the hospitals, as discussed in Section 6.1.4. As a result, an input utilising doctors, specialists and nursing staff could not be constructed and the number of nurses was used as the labour input. The numbers of nursing staff in each hospital for each year included both permanent staff, as well as agency staff.

The nursing data provided consisted of seven different nursing categories: nursing auxiliaries, pupil nursing assistants, pupil enrolled nurses, senior professional nurses, professional nurses, senior enrolled nurses and enrolled nurses. The different proportions of the types of nurses employed across the hospitals can be seen in Figure 6.5. It is clear from this figure that the extent to which different categories of nurses are employed, varies across hospitals. The smaller hospitals, particularly Hospital 1 and 8, appear to rely more heavily on enrolled nurses than many of the bigger hospitals. Only five hospitals make use of senior enrolled nurses. Pupil nursing assistants were only used in 2008 and 2009.

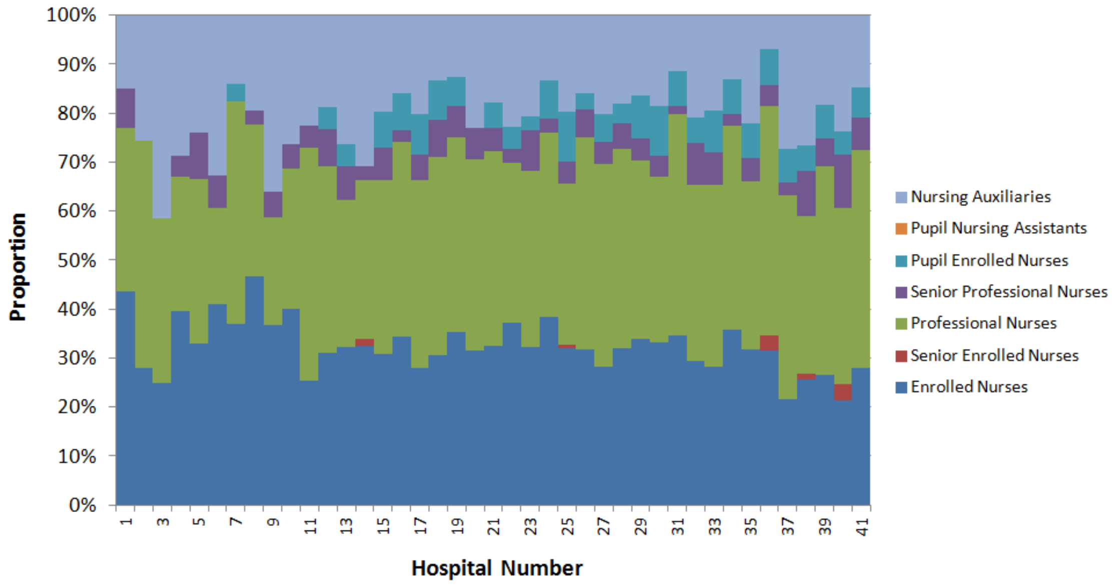

Figure 6.5: Proportion of each nursing category used in each hospital in 2011.

The relative salaries for each of these categories of nurses, which was assumed to be a proxy for experience and expertise, was used to construct a weighted average of the nursing staff, in order to create a single input. The number of nurses in each category was then multiplied by the respective calculated weight, to obtain the value for the adjusted number of nurses in each category. The adjusted number of nurses was summed across the categories, to obtain a single number for permanent nursing staff. Adjusting the numbers of nurses in this way allows for the different mixes of nursing staff across hospitals. For example, if two hospitals had the same number of nurses but the first hospital had a larger proportion of senior enrolled nurses and the second hospital had a larger proportion of pupil nurses, a straight aggregation does not capture the differences in seniority and expertise. The weighted aggregate ensures that the first hospital will have a larger number of nurses relative to the second hospital.

A series of simple checks were carried out to verify the adjustment. The first check carried out was to calculate the ratio of the actual number of nurses over the adjusted number of nurses, for each hospital, for each year. These ratios were checked for consistency over the four year period, for each hospital. No abnormalities were detected. The second check carried out was a comparison of the adjusted number of nurses across years, for each hospital. It appeared that the adjusted number of nurses across all hospitals increased dramatically from 2008 to 2009 and then decreased to around 2008 level in 2010. Upon further investigation it was found that the human resource data for the period September 2008 to August 2009 was doubled. To correct for this, the annual human resource figures were ratioed down to achieve a correct set of figures.

Calculating the adjusted number of agency nurses was more complex. The agency data were extracted from two different systems and had to be combined. In order to combine it, the hours worked and total amount paid was extracted for professional nurses, enrolled nurses and enrolled nursing assistants category. For the months where the two data sets overlapped, the most recent data set was used. Having established the hours worked and total amount paid for each category for every month, it was discovered that agency data were only available from mid-2008. As a result, the number of hours worked and the total amount paid per category were imputed for the first half of 2008.

The major difference between the full-time nursing data set and the agency staff data set is that the full-time nursing data set gives the number of people employed, whilst the agency data set gives the total number of hours worked. To combine these two data sets, the agency staff number of hours needed to be converted into a full-time equivalent number of employees. Before this was done, an identical procedure was used to adjust the agency staff number of hours to reflect differences in the mix of staff across hospitals. The adjusted number hours was divided through by 1760 to obtain the number of full-time equivalent agency staff. The figure of 1760 was arrived at by assuming agency staff work on average eight hours a shift for 20 shifts a month and work for 11 months of the year.

Figure 6.6 illustrates the number of agency staff, as well as the number of full time nursing staff, in each hospital in 2011. Hospital 14 makes the largest use of agency staff, approximately half of the staff working are agency staff. On the other hand, agency staff make up only 3% of the nursing staff in Hospitals 10 and 11. Whilst Hospital 41 has marginally more beds than Hospital 40, Hospital 40 uses 100 more nurses, most of which are full time. As expected, in general, the number of nurses increases as the number of operational beds increases.

Two checks were carried out on the adjusted and aggregated nursing data set. First, the number of nurses per operational bed was checked. The highest number of nurses per bed in 2011 was 1.79 for Hospital 40. Hospital 2 had the fewest nurses per operational

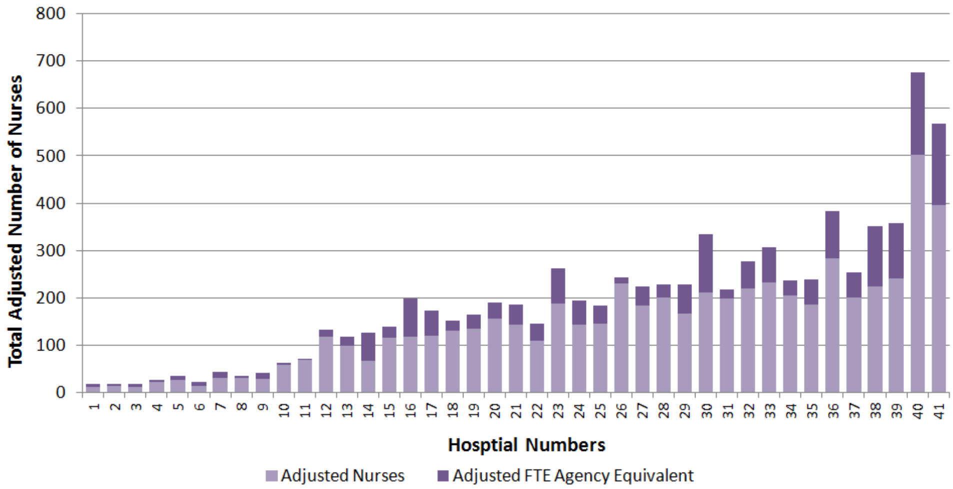

Figure 6.6: Number of agency staff and full-time nursing staff for each hospital in 2011.

bed in 2011 with a ratio of only 0.83 nurses per bed. Over the years, the ratio for most hospitals remained fairly constant. Although the number of nurses and number of operational beds is correlated, nurses is included as a separate input. The reason for including it is that nursing staff is an operational reality.

The second check carried out was to plot the number of nurses for each hospital for each year. This is shown in Figure 6.7. The number of nurses remains fairly constant over the four year period for the majority of the hospitals. This confirms that the adjustment to the 2009 data set was reasonable. Hospitals 33, 34 and 35 appear to experience a decrease in the number of nurses over the four year period. Whilst this appears odd, nothing unusual was found in the underlying data.

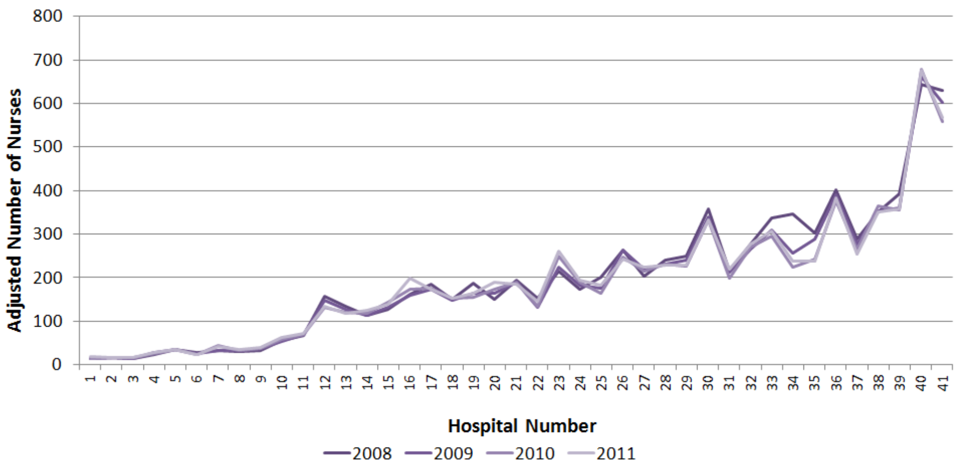

Figure 6.7: Adjusted number of nurses for the period 2008 to 2011.

Pharmacy Index

As discussed previously, the pharmacy input represents the goods that are required in the production process. Rather than use an absolute pharmacy billed amount, the ratio of the actual over the expected pharmacy expenditure was used. The advantage of using this ratio is that the absolute level of pharmacy expenditure does not affect the calculated efficiency scores for a hospital. This is particularly important when analysing efficiency over time, as an appropriate inflation adjustment is necessary if the absolute amount is used.

The ratio of actual expenditure to expected expenditure provides an indication as to whether a hospital is using more pharmaceuticals than expected. Analysing this ratio alone means that hospitals will appear inefficient if more pharmaceutical goods are used than expected.

This ratio was calculated by dividing an index for the actual pharmaceutical expenditure by an index for the expected pharmaceutical expenditure per case. The index for the actual pharmaceutical expenditure was calculated by first determining the average pharmaceutical spend per case, per hospital. This was then divided by the average pharmaceutical spend per case, over all hospitals. The index for the expected pharmaceutical expenditure was calculated using a factor provided in the data set, which indicates the expected relative expenditure per case. A factor for each hospital was calculated by averaging over all cases. An overall factor for the hospital group was calculated by multiplying the average factor for each hospital by the number of cases for that hospital and aggregating over all hospitals. This sum was then divided by the total number of cases across all the hospitals, to obtain an average factor for the hospital group. The expected index was calculated by dividing the factor for each hospital by the average factor for the hospital group. This process was carried out for each of the four years. These calculations were checked by ensuring that both the numerator and the denominator average to one across all hospitals.

The interpretation of this input is intuitive. If a hospital is utilising more pharmacy goods than expected, the pharmacy index will be greater than one. Similarly, the index will be less than one is the hospital is spending less on pharmaceutical goods than expected. The majority of the hospitals have a ratio of less than one. Five hospitals have a ratio greater than 1.1 which indicates that these hospitals spend 10% more on pharmaceutical goods than expected. These ratios can be seen in Table 6.1.

The five hospitals with ratios greater than 1.1 are all large hospitals. Figure 6.8 illustrates the pharmacy index for 2008 and 2011. The hospitals are ordered by the number of beds, from smallest to largest, and it appears as if the pharmacy index increases with the number of beds. In other words, on average the bigger hospitals tend to have a higher pharmacy index than the smaller hospitals. Over the four year period, most

Table 6.1: Hospitals with actual over expected ratios of over 1.1 in 2011.

Hospital No.	2011 Ratio
40	1.25
39	1.21
34	1.16
35	1.14
27	1.13

hospitals experienced a decrease in the pharmacy index. This decrease could be a result of prescribing and using more generic drugs when treating patients, or simply as a result of tighter controls and more efficient use of pharmaceuticals when treating cases. It is important to note that doctors control pharmacy usage through prescriptions. Therefore, there is potentially more variation present in pharmacy goods than in other inputs.

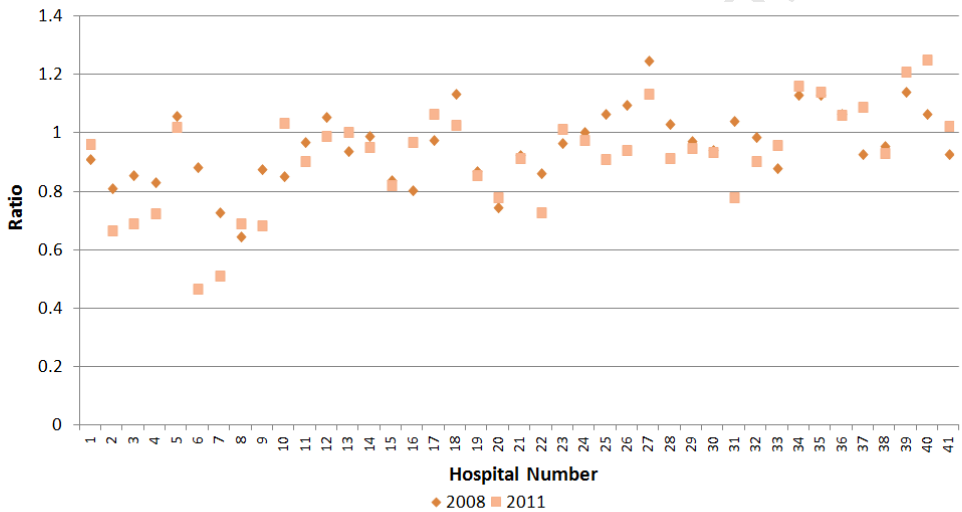

Figure 6.8: Pharmacy index for 2008 and 2011.

Theatre Minutes

Minutes for both major theatre procedures and minor theatre procedures were available in the data set. These were combined to create a single output. However, the two cannot simply be aggregated, as minor theatre procedure minutes will use significantly fewer resources than major theatre procedure minutes. If a simple aggregation was used, and two hospitals produced the same level of theatre minutes, but one through many major procedures and the other through minor procedures, the hospital with many major procedures is at risk of being identified as inefficient, as it will use substantially more resources to produce the same level of theatre minutes. Based on discussions with the executives of the private hospital group, it was decided to multiply major theatre minutes by a factor of 2.8, in order to make them comparable to minor theatre minutes. That is one major theatre minute is equivalent to 2.8 minor theatre minutes.

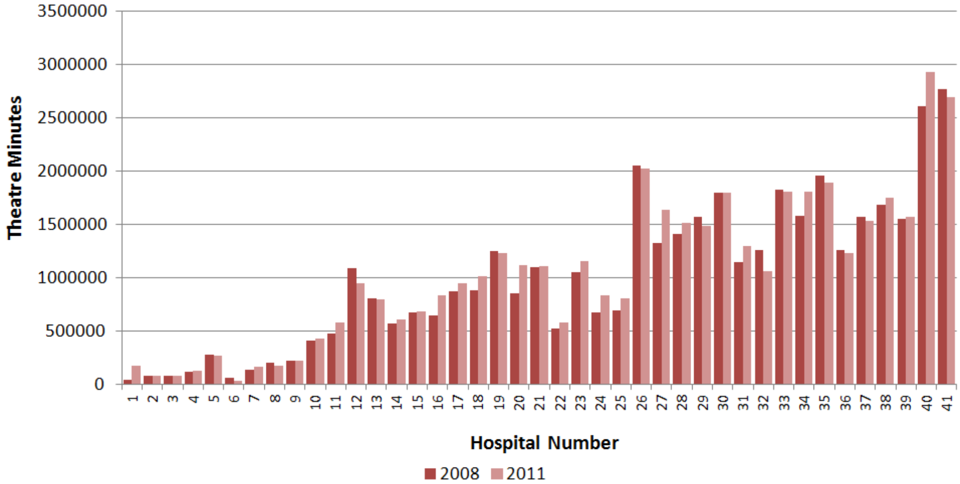

Figure 6.9: Total theatre minutes for 2008 and 2011.

Figure 6.9 illustrates the total number of theatre minutes for each hospital for 2008 and 2011. The hospitals are ordered by number of beds and it is clear that the bigger the hospital, the greater the number of theatre minutes. Overall theatre minutes do not appear to change significantly year to year. The only major change that occurred over the four year period was for Hospital 1 which increased the number of theatre minutes by approximately three times. The largest absolute change was approximately 320 000 theatre minutes for Hospital 40, which represents a 12.5% increase. To check that these changes in the total number of theatre minutes were not unreasonable, the number of theatre minutes per surgical case was analysed, see Figure 6.10.

The number of theatre minutes per case remain almost the same in 2008 and 2011 for the majority of the hospitals. For hospitals with more than 50 operational beds, that is, Hospital 11 to Hospital 41, theatre minutes per surgical case vary between 500 minutes at the top end and 220 minutes at the lower end. The variation between hospitals can be explained by three different factors. The first comprises the differences in the severity of cases that undergo surgery, with the more severe cases increasing the average time in theatre. The second comprises the proportion of cases that are not classified as surgical cases, but still undergo some surgical procedure, thereby increasing the number of theatre minutes, while number of surgical cases remains the same. For example ceasarian sections will result in theatre minutes but are classified under maternity cases and not surgical cases. A hospital with a large number of these cases will have a higher number of theatre minutes per surgical case, as only cases classified as surgical are included in the calculation. The third comprises of the variation in the types of doctors across hospitals.

Hospitals with fewer than 50 beds, Hospital 1 to Hospital 10, have a more volatile number of theatre minutes per surgical case. Hospital 4 has over 900 theatre minutes per surgical case. When analysing the number of theatre minutes across all cases, Hospital 4 falls into line with the rest of the hospitals. This is a possible indication that Hospital 4 has a very large proportion of cases that are not classified as surgical, but undergo minor surgical procedures relative to other hospitals. Whilst Hospital 1 experienced an increase in the number of theatre minutes from 2008 to 2011, the number of theatre minutes per surgical case decreased over this period. This is likely to be as a result of treating five times more surgical cases in 2011 as compared to 2008.

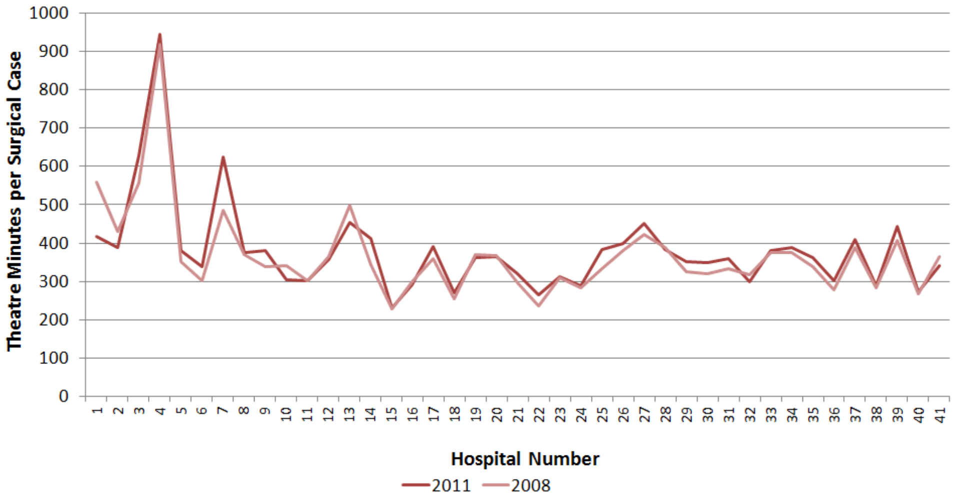

Figure 6.10: Total theatre minutes per surgical case for 2008 and 2011.

Number of Admissions

Eight broad admission categories remained after cleaning the data, namely Caesarean Section, Normal Vaginal Delivery (NVD), General Medical Admit, General Surgical Admit, Medical Day Case, Surgical Day Case, Medical Ambulatory and Surgical Ambulatory. These admissions were aggregated to determine the number of admissions per hospital per year.

The number of admissions is illustrated in Figure 6.11. The majority of hospitals experienced a very small change in the number of cases. The two exceptions to this are Hospital 1 and Hospital 7. Both these hospitals experience an approximately 50% increase in the number of admissions over the four year period.

It is clear from Figure 6.11, that the number of admissions increases with the size of

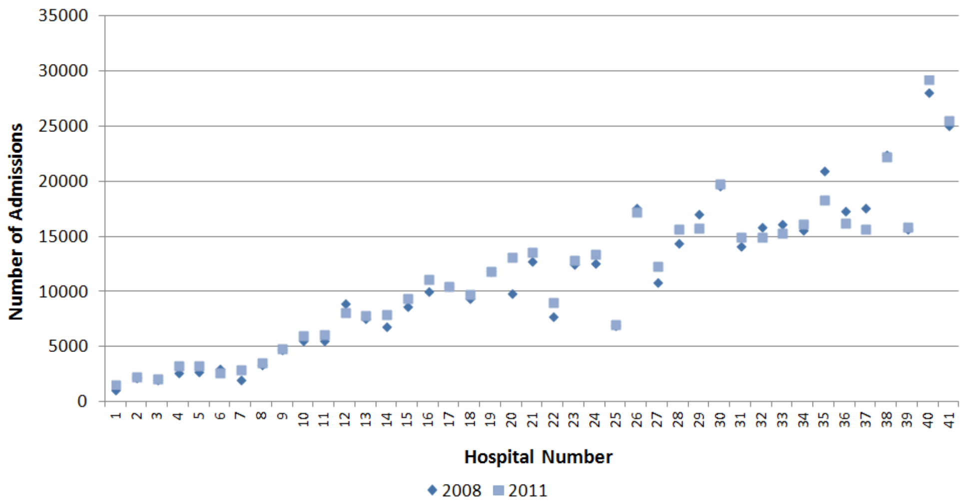

Figure 6.11: The number of admissions for 2008 and 2011.

the hospital, as is expected. However, there exist a few anomalies; of particular concern is Hospital 25. In spite of having 17 more operational beds in 2011 than Hospital 24, there are roughly half as many admissions. However, the occupancy rate for Hospital 25, as seen in Figure 6.2, is not out of line with the occupancy rates of the other hospitals. This suggests that the length of stay for each admission must be significantly longer than any of the other hospitals. This is not unlikely, given that Hospital 25 has a private HIV clinic and is likely to have a higher proportion of HIV-positive patients, with longer stays, in comparison to the other hospitals. Hospital 39 also appears to have fewer cases than hospitals of similar size, although the difference is not as extreme as that of Hospital 25. Hospital 40 and 41 have the same number of operational beds, however Hospital 40 treats approximately 4000 more patients a year in comparison to Hospital 41. There are no obvious possible reasons for these differences.

6.2.3 Adjusting for Case Mix

As discussed in Chapter 5, there are a number of different methods available to adjust for differences in case mix when using a DEA model. The four outputs used to analyse the effects of including a case-mix adjustment are:

- Unadjusted number of admissions
- Disaggregated number of admissions
- Case-mix adjustment factor as an additional variable
- Case-mix adjusted admissions

Each of these is described in detail below, except for unadjusted number of admissions, which has already been discussed in the data section directly above, as well as the construction of the case-mix adjustment factors.

Disaggregated Number of Admissions

Rather than use a case-mix grouper to disaggregate the admission data, as suggested by Sherman (1984), the admissions are disaggregated using the broad admission categories mentioned above. To avoid increasing the number of outputs too much, and reducing the power of the model, as discussed in Section 4.2, these eight broad admission categories were grouped into five larger categories. A maternity category was created using the Caesarian Section and NVD admissions (the split between caesarian sections and NVDs does not vary substantially across hospitals), Medical Day Cases and Surgical Day Cases were grouped into a Day Cases category; and Medical Ambulatory and Surgical Ambulatory were grouped into an Ambulatory category. General Surgical Admits and General Medical Admits formed their own categories. Figure 6.12 shows the proportion of total cases that fall within each of the five categories for each hospital in 2011.

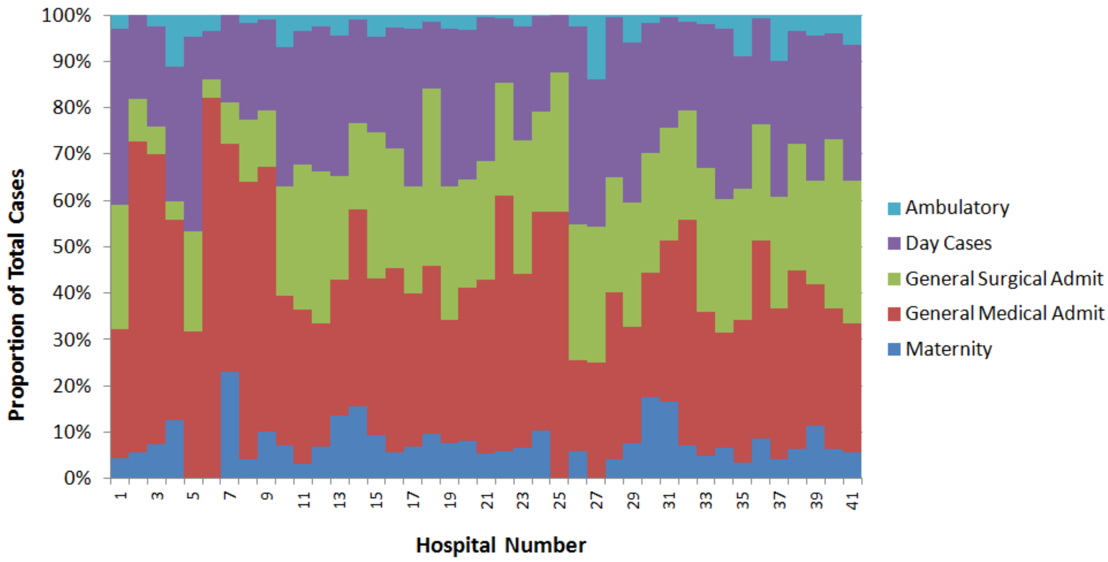

Figure 6.12: Disaggregated Admissions for 2011.

It is clear that the mix of cases varies across hospitals. Categories vary from a minimum of 0%, for example the Maternity category in Hospitals 5, 25 and 27, to a maximum of 82%, which is the proportion of General Medical Admissions for Hospital 6. The Ambulatory category is consistently a small proportion of admissions.

Constructing a Case-Mix Adjustment Factor

The method proposed by Fetter et al. (1980) was used to calculate case-mix adjustment factors for each hospital, for each year of the study. However, unlike Fetter et al. (1980), the billed amount rather than the length of stay is used as an indication of resource consumption. The reason for this is that resource consumption may vary significantly over the same length of stay. For example, one day in an Intensive Care Unit consumes significantly more resources than one day in a general ward. In order to capture this variation in resource consumption, the total billed amount was used rather than the length of stay. There are two disadvantages of using billed amount. Firstly, the billed amount is influenced by both patient-related characteristics as well as supply-side factors, such as doctors' prescribing and referral behaviour. As a result, the total billed amount does not exclusively capture the severity of illness or the total resources used because of patient characteristics, rather it is confounded by supply-side factors. The second disadvantage is that costs per procedure may vary across hospitals and geographic locations. However, this is not an issue for this analysis since the hospital group charges the same tariffs for procedures across all hospitals.

The following formula was used (Fetter et al., 1980):

$$CMAF_i = \frac{\sum_j A_j p_{ij}}{\sum_j A_j P_j} \text{ for } j = 1, 2, 3, \ldots \tag{6.1}$$

where

a_{ij} = the average billed amount in the j^{th} clinical group for hospital i,
p_{ij} = the proportion of hospital i's cases in the j^{th} clinical group,
P_j = the proportion of all hospital cases in clinical group j,
A_j = the average billed amount across all hospitals for the j^{th} clinical group,

A case-mix factor of less than one indicates that a hospital is treating relatively fewer severe cases in comparison to the average hospital. Similarly, a hospital with a case-mix factor of greater than one suggests that that hospital is treating relatively more severe cases in comparison to the average hospital.

In order to determine whether standardisation across the clinical groups is appropriate, the interaction component should be calculated for each of the hospitals (Fetter et al., 1980). The interaction component is calculated using the following formula:

$$\text{Interaction Component} = \sum_j (a_{ij} - A_j)(p_{ij} - P_j) \text{ for } j = 1, 2, 3, \ldots \tag{6.2}$$

As discussed in Section 5.6, the sign and magnitude of the interaction component provides insight into the extent to which both case mix and individual hospital-related factors are jointly accountable for the difference between the average billed amount per case for a hospital and the average billed amount per case across all hospitals. A large interaction component is an indication that case mix varies significantly across hospitals or that utilisation of resources varies across hospitals. If there is a large positive interaction component, standardisation may be misleading.

As per the international literature, case-mix adjustment factors were first calculated using DRGs as the clinical groups, using the formula above. The DRGs used in this analysis were constructed by the hospital group and there are a total of 1071 distinct groups. The calculated case-mix adjustment factors differ very little across years for each hospital. The calculated case-mix adjustment factors for each hospital for 2008 to 2011 are displayed in Figure 6.13. It is clear that case-mix adjustment factors do not vary substantially across the four years.

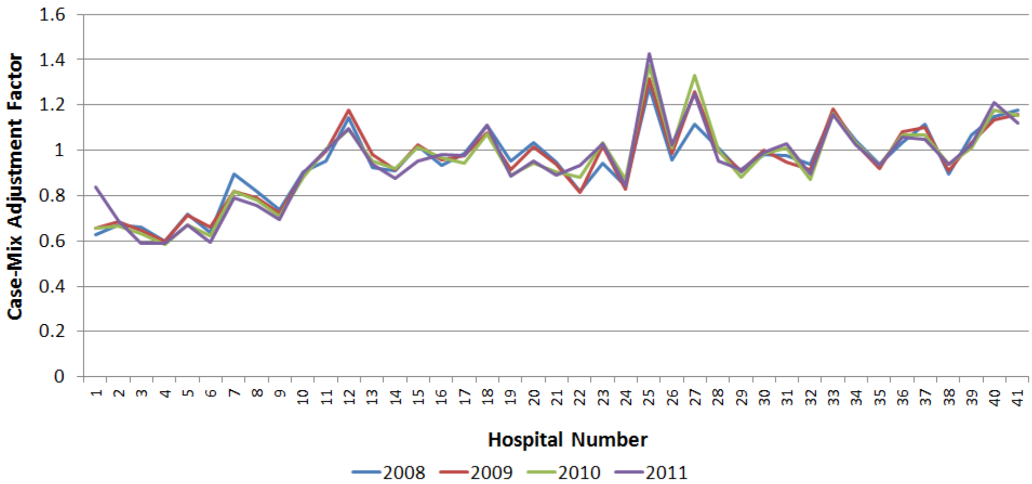

Figure 6.13: DRG case-mix adjustment factors for 2008 to 2011.

Since case-mix adjustment factors are very similar across years, interaction components were calculated for each of the hospitals in 2011 only. The interaction components varied across hospitals in terms of sign but are all relatively small. This confirms that standardisation across hospitals and DRGs was appropriate and is not misleading.

For interest, the interaction component was calculated for those hospitals that were excluded at the start of the study and not included in the final analysis. The interaction component for the specialist hospital was positive and very large in comparison to the other interaction factors; therefore, standardisation across this hospital would not be appropriate. This result further supports the exclusion of this hospital from the final analysis.

As a result of the well-documented heterogeneity of DRGs, the refinement of the calculated case-mix adjustment factors was investigated. Age and gender are two well-known indicators of severity and recovery period. For example, the risks of a mastectomy are far higher for patients over the age of 65, in comparison to the risks faced by a 30 year old. The idea was to disaggregate existing DRGs into four different groups: males under

65, males over 65, females under 65 and females over 65. Unfortunately, data become too scanty in many of the groups to disaggregate at the DRG level. Instead, Basic-Diagnostic-Related Groups (BDRGs) had to be used. BDRGs are broader than DRGs and do not account for differences in severity of illness. For example, the BDRG for spinal procedures has three DRGs associated with it, namely, spinal procedures without complications, spinal procedures with complications and spinal procedures with major complications. There are 365 different BDRGs and 1071 different DRGs. For each BDRG, four groups were created using age and gender. For the remainder of this dissertation BDRG refers to BDRGs that have been divided into smaller groups using age and gender.

The calculated BDRG case-mix adjustment factors, like the DRG case-mix adjustment factors, are very similar across all four years of the analysis. A comparison of the BDRG and the DRG case-mix adjustment factors for 2011 is shown in Figure 6.14. For the majority of the hospitals, there is almost no difference between the BDRG and DRG case-mix adjustment factors. The correlation coefficient of 0.975 confirms that there is a strong linear relationship between the two series of case-mix adjustment factors. There is only one hospital for which there is a noticeable difference between the two factors, namely Hospital 25. The DRG case-mix adjustment factor is larger than the BDRG case-mix adjustment factor for Hospital 25. This may be an indication of DRGs capturing more of the variation present in billed amounts, in comparison to the constructed BDRGs. Theoretically, depending on the mix of cases, DRGs may be better at explaining variation in some cases, whilst BDRGs may be better in other cases.

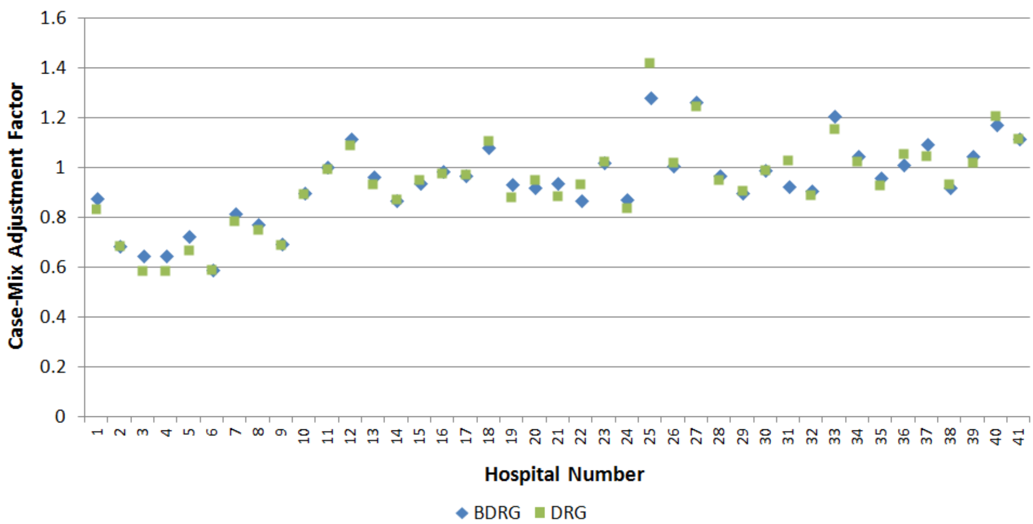

Figure 6.14: A comparison of BDRG and DRG case-mix adjustment factors for 2011.

It is perhaps important to digress at this point, and note the differences between the case-mix adjustment factors used for alternative reimbursement and those factors used for efficiency studies. Under alternative reimbursement it is essential that the case-mix adjustment factors for each group are accurate, such that hospitals are paid a fair amount for each case and that cross-subsidisation among cases is minimised. Unlike alternative reimbursement, case-mix factors used in efficiency studies do not have to be accurate at an individual diagnosis group level. Rather, it is important that the overall factor for the hospital accurately represents the mix of cases relative to the mix of cases experienced by other hospitals. For this reason, the case-mix adjustment factors are analysed and compared on the hospital level only.

The process of refinement was taken one step further by adjusting the data for outliers.

Upon closer inspection of the groups which were aggregated over hospitals to construct the case-mix adjustment factors, it is clear that the distribution of billed amounts is skewed and in some groups there exist outliers. For example consider the histogram of billed amounts for 'Thyroid, Parathroid and Thyroglossal Procedures for females under the age of 65', shown in Figure 6.15. It is clearly skewed to the right and is not symmetrically distributed. Furthermore, there appear to be extreme outliers in the right tail of the distribution.

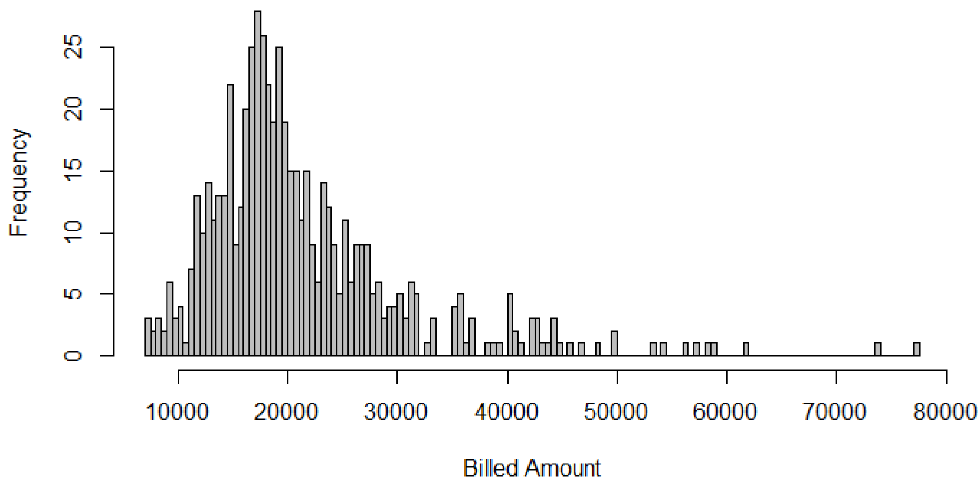

Figure 6.15: Histogram of billed amounts for 'Thyroid, Parathroid and Thyroglossal Procedures for females under the age of 65' in 2011.

It is evident from Figure 6.15 that using the average billed amount per case would not necessarily provide an accurate indication of the expected resource consumption for a case in the group. For this reason, it is necessary to examine outliers to determine the extent to which they influence the mean billed amount.

Only groups with more than 100 cases in the group were analysed for outliers. In groups with fewer than 100 cases there is a risk that there are too few cases to accurately analyse the true distribution of billed amounts is distorted. There are a total of 574 groups with more than 100 cases that were analysed for outliers.

The adjusted boxplot method for skewed distributions introduced by Vanderviere and Huber (2004) was used to detect outliers. The original boxplot method developed by (Tukey, 1977) assumes normality and is inappropriate when the underlying distribution is skewed, as too many points are identified as outliers (Vanderviere and Huber, 2004).

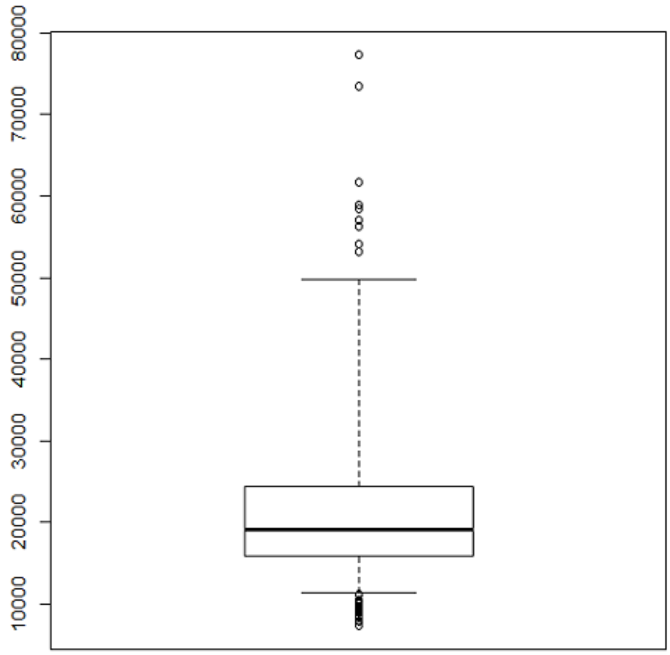

Figure 6.16: Adjusted boxplot for billed amounts for 'Thyroid, Parathroid and Thyroglossal Procedures for females under the age of 65' in 2011.

The adjusted boxplot for 'Thyroid, Parathroid and Thyroglossal Procedures for females under the age of 65' is shown in Figure 6.16. As suspected, that there are a number of outliers. For this particular group, there are outliers in both the left and right tails of the distribution of billed amounts. These identified outliers were removed from this group and a new average billed amount per case was calculated and used in the construction of the case-mix adjustment factors.

This method was carried out on all 574 groups with more than 100 observations. Outliers were identified and removed from 560 out of the 574 groups. The proportion of outliers identified within a group ranged from 0% to 25.2%. The group for which 25.2% of cases were identified as outliers was 'Other female reproductive system procedures for females over the age of 65'. This group is a catch-all group and encompasses all cases that do not fall into the other female reproductive groups. As a result, the individual cases are varied and the billed amounts range from under R5 000 to over R150 000. Therefore, it is not unusual, or unexpected, that so many cases were identified as outliers. A high proportion of outliers was also detected in the group 'Other female reproductive system procedures for females under the age of 65', where 18.4% of the cases were detected as outliers. The highest number of outliers detected was 722 in 'Cesarean Deliveries for females under the age of 65', however these outliers were only 2.9% of all the cases in this group.

Since the mix of cases across years appears to be stable, outliers were detected for 2011 only. After removing the outliers from the 574 groups, new average billed amounts per case were calculated for each of the groups. The same method described previously was used to calculated the case-mix adjustment factors. However, rather than using average billed amount per case, per group, per year, the average billed amount per group for 2011 was used across all four years. These new averages were used in conjunction with the actual number of cases in each year to construct an adjustment factor for each hospital.

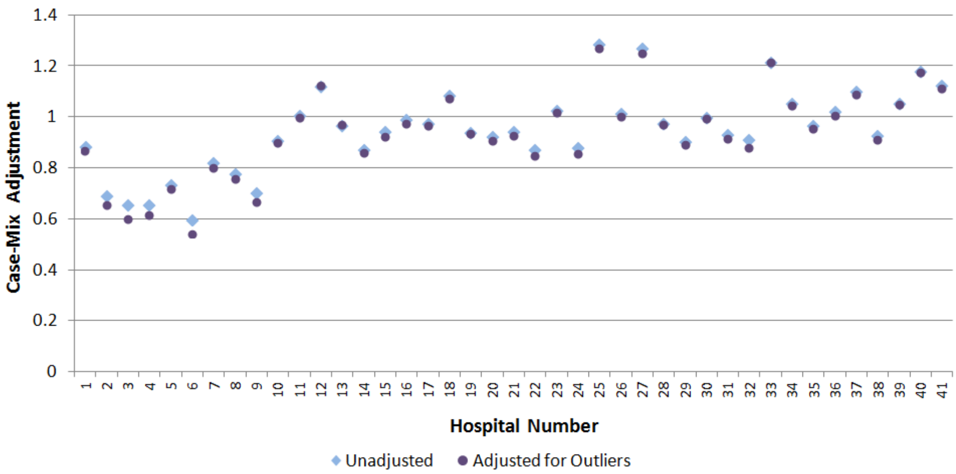

Figure 6.17: A comparison of the case-mix adjustment factors adjusted for outliers and the unadjusted values in 2011.

The case-mix adjustment factors adjusted for outliers were compared with the unadjusted case-mix adjustment factors and very little difference was noted. Figure 6.17 displays the two sets of case-mix adjusted factors, the one adjusted for outliers and the other not. The factors are almost identical in value except for a few small hospitals, where the factors adjusted for outliers are slightly lower than the unadjusted values.

At this point, it is perhaps useful to reiterate that the accuracy of the individual groups is not critical. Instead, the accuracy of the aggregated groups and the factor for the hospital as a whole is important. Since there is very little difference between the case-mix adjustment factors adjusted for outliers and those that are unadjusted, and the accuracy of the individual groups is not critical, it was decided to leave the case-mix factors unadjusted for outliers. Cases identified as outliers can, and do, occur in a hospital environment and it can be argued that excluding these cases is unrealistic.

For the purposes of measuring case mix in this dissertation, both the DRG case-mix adjustment factors and BDRG case-mix adjustment factors are used in final models in Chapter 7.

Incorporating Case-Mix Adjustment Factors

There are two methods available to incorporate case-mix adjustment factors into outputs, as discussed in Sections 5.5 and 5.6.

- Case-Mix Adjustment Factors as an Additional Output

  The DRG case-mix adjustment factors are used as an additional output. As previously mentioned, these factors remain relatively constant across the four year period, as seen in Figure 6.13. Fourteen hospitals in 2011 have case-mix adjustment factors over one.

- Case-Mix Adjusted Admissions

  DRG Case-Mix Adjusted Admissions
  The adjusted number of admissions was calculated by multiplying the unadjusted number of admissions by the DRG case-mix adjustment factor for each hospital.

  A comparison of case-mix adjusted and unadjusted admissions, as well as the adjustment factors for 2011, are shown in Figure 6.18. As expected, the smaller hospitals case-mix adjusted admissions are lower than the unadjusted admissions because the case-mix adjustment factors are less than one. The result of this is that small hospitals appear smaller after adjusting for case mix, and large hospitals appear even larger. The reason for the relationship between scale and the case-mix adjustment factors is that large hospitals are likely to have a number of different specialties, attracting very ill patients that require specialist treatment. Furthermore, large hospitals are likely to be better equipped and have the capacity to cope with a large number of complex cases, in comparison to smaller hospitals. This relationship is important to keep in mind when analysing the scale efficiency of a hospital, as the case-mix adjustment factor ultimately alters the scale of production by adjusting the number of cases treated.

  Hospital 25, despite having the highest case-mix adjustment factor of 1.43, still has relatively low adjusted admissions in comparison to hospitals of a similar size. Hospitals 3, 4, and 6 experienced a roughly 40% decrease in admissions, to obtain adjusted admissions. The two largest hospitals, Hospitals 40 and 41, increased by approximately 6000 and 3000 admissions, a change of 21% and 12% respectively.

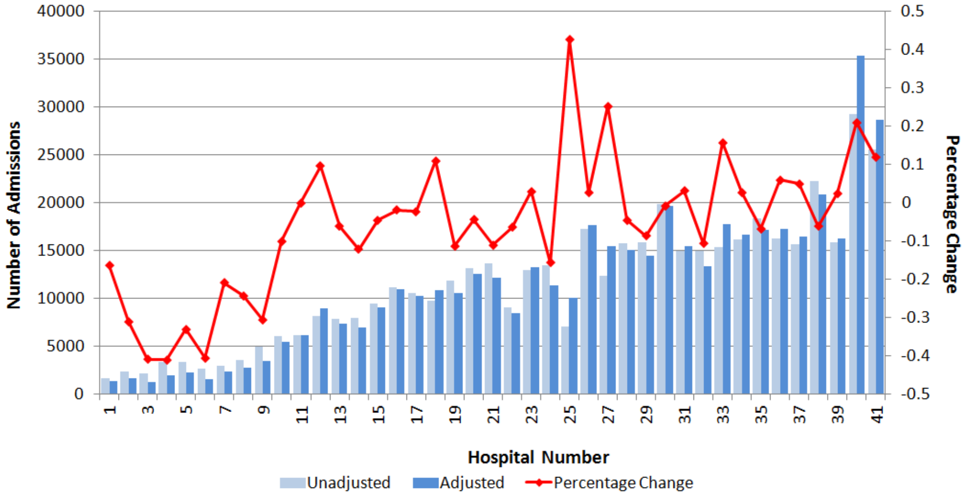

Figure 6.18: A comparison of DRG case-mix adjusted and unadjusted admissions for 2011.

BDRG Case-Mix Adjusted Admissions

As with the DRG case-mix adjusted admissions, the BDRG adjusted number of admissions was calculated by multiplying the unadjusted number of admissions by the BDRG case-mix adjustment factor for each hospital. A similar figure to Figure 6.18 for BDRG case-mix adjusted admissions was constructed. As expected, there is very little difference between the two figures as two sets of case-mix adjustment factors are almost identical. The same relationship between the case-mix adjustment factor and size of the hospital was present.

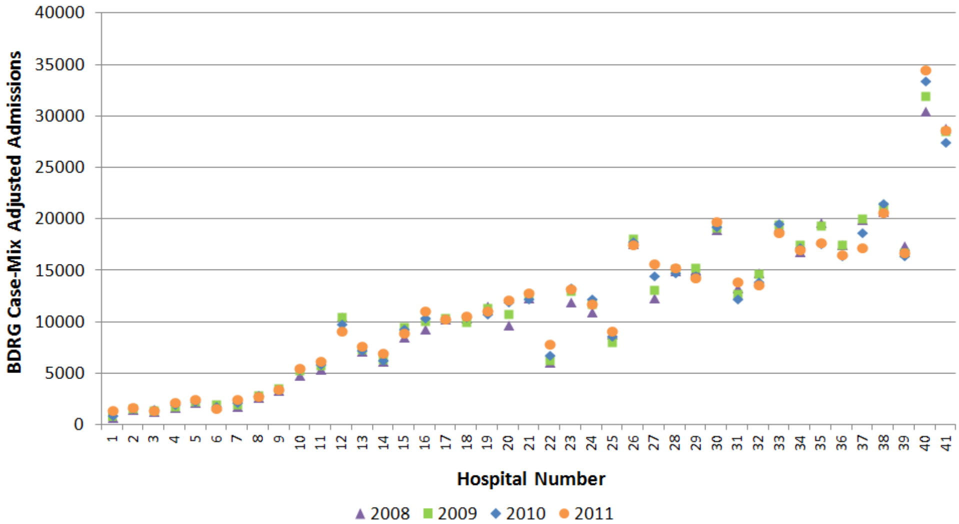

Figure 6.19: BDRG case-mix adjusted admissions for 2008 to 2011.

Figure 6.19 illustrates the BDRG case-mix adjusted number of admissions for the four year period of the analysis. Year to year, the adjusted number of admissions hardly changes for the majority of the hospitals. The one obvious exception to this is Hospital 40 which shows a consistently increasing number of BDRG case-mix adjusted cases over the years. This change can be attributed to an increasingly severe mix of cases. Even after adjusting for case mix, Hospitals 22, 25 and 39 still appear to treat relatively few cases in comparison to hospitals of similar sizes. A very similar pattern was visible when analysing the DRG case-mix adjusted admissions over the four years.

6.2.4 Orientation

The choice of orientation should be driven by the objectives of the analysis and any management constraints (Nguyen and Coelli, 2009). O'Neill et al. (2008) note that the focus of most hospitals is on reducing and controlling costs, rather than influencing the demand for health care. For this reason, an input orientation was chosen. This is consistent with much of the hospital efficiency literature (O'Neill et al., 2008).

6.2.5 Returns to Scale

A model can be specified as either having CRS or VRS. Running the models under CRS is inappropriate, as it assumes that all the units are operating at the optimal size (Nguyen and Coelli, 2009). Since the hospitals vary significantly when it comes to size and the optimal size of a hospital is unknown, this is an inappropriate assumption. The obvious choice is to assume VRS to avoid confounding efficiency scores.

Sherman and Zhu (2006) recommend running models with both CRS and VRS in order to determine and analyse the presence of scale efficiencies. However, all other efficiencies are analysed under VRS.

6.2.6 Choice of Software

Microsoft Office Excel 2007 and *Microsoft Visual Basic (Version 6.5)* were used to extract the required data from the data set and construct the chosen inputs and outputs. These programmes were also used for the initial data analysis.

RStudio (version 0.95.265), developed by the R Foundation for Statistical Computing (2012), was used to run the DEA models. In particular the Benchmarking package, developed by P. Bogetoft & L. Otto (2011), was used for the DEA frontier analysis and the FEAR package, developed by P. Wilson (2010), was used to analyse efficiency over time and to construct the Malmquist Total Factor Productivity index. The Robustbase package, developed by P. Rousseeuw & C Croux (2012), was used to determine outliers.

Chapter 7

Results

7.1 Description of the Models

Five models were constructed and examined. The outputs of these five models varied:

- **DRG model**: one output, namely Diagnosis-Related Group (DRG) case-mix adjusted number of admissions;
- **Standard model**: one output, namely unadjusted admissions;
- **Disaggregated model**: five outputs, namely maternity cases, day cases, ambulatory cases, general medical admits and general surgical admits;
- **Additional output model**: two outputs, namely unadjusted admissions and the DRG case-mix adjustment factors; and
- **BDRG model**: one output, namely Basic-Diagnosis-Related Group (BDRG) case-mix adjusted number of admissions, where the BDRG case-mix adjustment factors allow for age and gender as well.

The five models had three inputs in common, namely, operational beds, the adjusted number of nurses and a pharmacy index.

Three major comparisons are carried out using these five models. Figure 7.1 illustrates these comparisons.

In order to contextualise results, detailed results of the DRG model are discussed in isolation first. This is followed by comparison of the DRG model and the standard model. The aim of this comparison is to determine the impact that a case-mix adjustment has on efficiency scores. After ascertaining the impact adjusting for case mix has on efficiency scores, a comparison of the different methods of adjusting for case mix is carried out. A refinement of the case-mix adjustment factor was also considered and the results of the DRG and the BDRG model are compared in order to determine whether the refinement has a material impact on efficiency scores. This chapter concludes with an analysis of individual hospitals which have interesting or odd results.

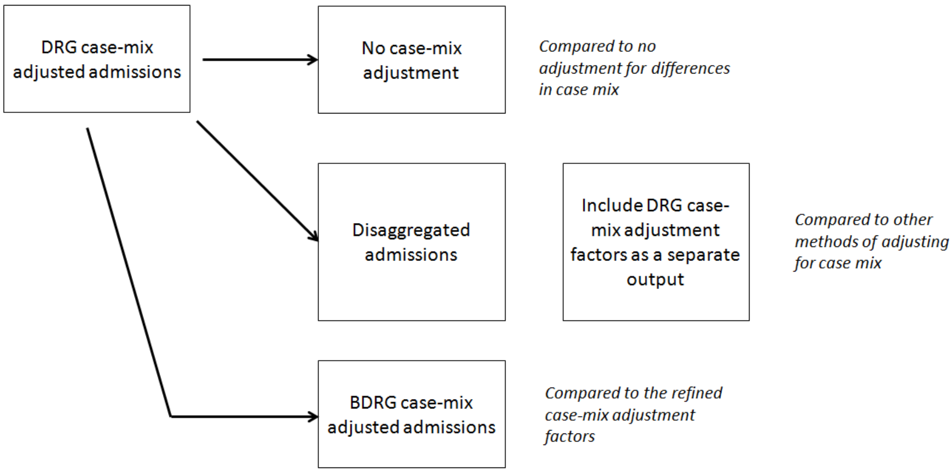

Figure 7.1: Illustration of the final models and comparisons.

It is important to make a point on interpreting results at this juncture: as mentioned in Chapter 4, DEA is unable to account for random variation, instead the random variation is classified as an inefficiency. Therefore, when interpreting results over time, slight differences may be a result of random variations and do not necessarily indicate a change in efficiency. For this reason, the focus of the results is the change over the medium term from 2008 to 2011, rather than annual changes. Large changes experienced by individual hospitals are also investigated.

7.2 A Summary of the DRG Model

A summary of the results of this model are shown in Table 7.1.

Table 7.1: Summary of efficiency scores for the DRG model.

Year	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient
2008	0.91	31.71%
2009	0.93	39.02%
2010	0.93	34.15%
2011	0.92	29.27%

Average efficiency scores differ very little over the four year period. In 2011 the average efficiency score was 92%, this means that on average resource savings of 8% can be made. Whilst the average efficiency score increases slightly over the four year period, the proportion of hospitals which are fully efficient[1] decreases slightly. Over the four

[1]A hospital is defined 'as fully efficient', or DEA efficient (see Section 4.6.3), if it has an efficiency score of one and there are no slacks.

year period, roughly 30% of the hospitals were classified as fully efficient. The average efficiency scores indicate a minor improvement in the average efficiency of the group. However, there are fewer hospitals which are operating efficiently. This indicates that on average hospitals have improved efficiency, but fewer are fully efficient. It is clear that analysing either of these measures in isolation would provide a distorted picture of efficiency. As a result, it is necessary to evaluate the proportion of hospitals that are fully efficient in conjunction with the average efficiency score for the group, as well as the Malmquist Total Factor Productivity index to better analyse changes in productivity and efficiency over time.

Malmquist Total Factor Productivity (TFP) indices were calculated for each of the hospitals using all the data for the four year period. As discussed in Section 4.7, the TFP index can be disaggregated into an index representing technological change and an index for technical efficiency change. The index for technological change represents a shift in the production frontier, whilst the index for technical efficiency change measures the changes in the distance from this frontier. The index for technological change can be disaggregated into an index for pure technological change and an index for the change in the scale of technology. Similarly the index for pure technical efficiency change can be disaggregated into an index for pure technical efficiency change and an index for the change in scale efficiency. The technical efficiency change index can be equated to changes in the technical efficiency scores under CRS, whilst the pure technical efficiency change index is associated with changes in technical efficiency under VRS. Pure technological change and the scale of technological change is not investigated in any detail.

The overall percentage changes in productivity, technology, technical efficiency, pure technical efficiency and scale efficiency are listed in Table 7.2. Overall, productivity increased by 3.4%, which was largely driven by the 4.68% increase in technical efficiency. Both components of technical efficiency increased from 2008 to 2011.

Table 7.2: Changes in the components of the Malmquist Total Factor Productivity Index.

Index	% Change from 2008 to 2011
TFP	3.40%
Technological Change	-1.06%
Technical Efficiency Change	4.68%
Pure Technical Efficiency Change	1.09%
Scale Efficiency Change	3.66%

The five hospitals with the largest increase in pure technical efficiency from 2008 to 2011, and the five hospitals with the largest decrease in pure technical efficiency, were identified using the Malmquist pure technical efficiency indices for 2008 to 2011. The efficiency scores for these hospitals are listed in Table 7.3.

Table 7.3: The largest overall increases and decreases over the four year period for pure technical efficiency.

Largest Overall Increase in Pure Technical Efficiency			
Hospital Number	2008	2011	Increase
34	0.82	0.96	17.07%
6	0.87	1.00	14.94%
27	0.83	0.94	13.25%
25	0.68	0.77	13.24%
23	0.80	0.88	10.00%
Largest Overall Decrease in Pure Technical Efficiency			
Hospital Number	2008	2011	Decrease
14	0.91	0.76	16.48%
37	1.00	0.89	11.00%
39	0.73	0.67	8.23%
8	1.00	0.93	7.00%
15	0.95	0.89	6.32%

Hospital 34 experienced the largest increase in pure technical efficiency from 2008 to 2011. Hospital 6 is the only hospital that experienced an increase in efficiency, as well as being classified as fully efficient in 2011. Although Hospital 25 experienced a large increase in pure efficiency over this period, the actual efficiency score for 2011 is fairly low in comparison to the other hospitals.

Hospital 14 experienced the largest decrease in pure technical efficiency from 2008 to 2011. Interestingly, the efficiency scores remained fairly stable from 2008 to 2010 and then decreased suddenly in 2011. This sudden decrease is unexpected and warrants further investigation. Hospital 39 has relatively low efficiency scores to start with in 2008. The fact that it experienced a large decrease in efficiency over this period is concerning. Both Hospital 14 and Hospital 39 are investigated further in Section 7.6.

Individual hospitals with low relative efficiency scores were also examined. Five hospitals were identified with efficiency scores equal to, or less than, 0.8 in 2011. These hospitals are listed in Table 7.4. A cutoff value of 0.8 was chosen somewhat arbitrarily. The motivation for this cutoff is that it indicates that efficiency improvements of at least 20% can be made before a hospital is classified as fully efficient.

Only one of the five hospitals listed in Table 7.4, Hospital 25, shows an improvement in efficiency from 2008 to 2011; the other four hospitals show a worsening of efficiency.

Table 7.4: Hospitals with efficiency scores of 0.8, or lower, in 2011, for the DRG model.

Hospital	2008	2011
39	0.73	0.67
14	0.91	0.76
25	0.68	0.77
36	0.81	0.80
24	0.85	0.80

Hospital 39 has the lowest efficiency score in 2011, and the second lowest efficiency score in 2008.

Efficiency scores for all the hospitals for the period 2008 to 2011 can be found in Appendix A, Table A.1.

Seven hospitals were fully efficient in each of the four years of the analysis. These seven hospitals are: Hospital 1, Hospital 2, Hospital 3, Hospital 10, Hospital 30, Hospital 41 and Hospital 40. Not one of these seven hospitals had comparable hospitals, known as peers, and as a result were allocated efficiency scores of one. As explained in Chapter 4, these hospitals were classified as efficient as there was insufficient information available to suggest otherwise. The same applies to the 12 hospitals identified as fully efficient in 2011. This finding highlights one of the dangers of analysing relative efficiency scores in isolation. Figure 7.2 illustrates the ERS for each hospital.

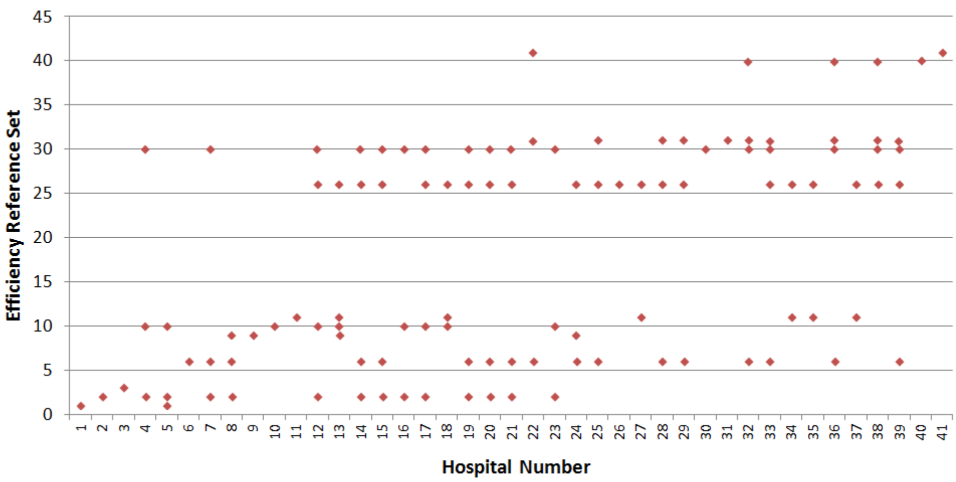

Figure 7.2: An illustration of the ERS's for each hospital in 2011.

It is clear from Figure 7.2 that for Hospital 1, Hospital 2, and Hospital 3, there are no other hospitals in their ERS except for themselves. Hospital 5 has three peer hospitals and the ERS consists of Hospital 1, Hospital 2, and Hospital 10. The number of times a hospital appears in an ERS is the horizontal count of the points across the figure. For example, Hospital 26 appears in 21 ERS's. Hospital 2, Hospital 6 and Hospital 30 also appear in a large number of ERS's.

The results highlighted in this section emphasise the importance of looking at a range of different measures. For example, it is clear that average efficiency scores for the group do not provide a reliable indication of the performance of all hospitals. Instead, average efficiency scores should be assessed in combination with other measures, such as the proportion of hospitals which are fully efficient and Malmquist TFP indices. A second example is hospitals with efficiency scores of one. Whilst these hospitals appear fully efficient when evaluating only the efficiency scores, many hospitals do not have peer hospitals. So, these hospitals are classified as efficient by default, rather than being truly efficient relative to their peers.

In summary, the change in average technical efficiency for all hospitals over the four year period is close to 5%, as identified by the Malmquist TFP index.

7.3 The Impact of Adjusting for Case Mix

The aim of this section is determine the impact that adjusting for case mix has on efficiency scores. Hospitals with a heavier case mix, in other words a case-mix adjustment factor of over one, are expected to appear more efficient under the DRG model, in comparison to the standard model. Similarly, those hospitals with a lighter case mix are expected to appear less efficient under the DRG model, in comparison to the standard model.

Summary results of the standard model and DRG model are shown in Table 7.5. Efficiency scores for all hospitals over the four year period of the analysis for the standard model can be found in Appendix A, Table A.2.

Table 7.5: Summary of efficiency scores for the standard model and the DRG model

	Standard Model		DRG Model	
Year	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient
2008	0.91	24.39%	0.91	31.71%
2009	0.92	31.71%	0.93	39.02%
2010	0.91	29.27%	0.93	34.15%
2011	0.90	36.59%	0.92	29.27%

As seen in Table 7.5, the average efficiency scores of the standard model are very similar to those of the DRG model. However, rather than the average efficiency scores increasing marginally over the four year period, as in the DRG model, the average efficiency scores of the standard model decrease marginally. The proportion of hospitals which are fully efficient is higher in the DRG model for each year, except for 2011, where there are three more hospitals which are fully efficient under the standard model in comparison to the DRG model.

To better understand this dynamic, the efficiency scores for each of the models are compared for 2011, as shown in Figure 7.3.

It is clear in Figure 7.3 that adjusting for case mix has a large impact on individual efficiency scores for hospitals. Figure 7.3 is a good illustration of the dangers of not adjusting for case mix when case mix is not constant across hospitals.

There is only one hospital which was identified as relatively inefficient under the standard model and fully efficient under the DRG model, namely Hospital 11. Hospital 11 provides an interesting example of the manner in which DEA evaluates decision-making units (DMUs) relative to their peers. The case mix adjustment factor for Hospital 11 is very close to one in 2011. As a result, the case-mix adjusted outputs and the unadjusted outputs are almost identical. Under the standard model, Hospital 11 uses relatively more inputs to produce a single output in comparison to its peers. For this reason, Hospital 11 is inefficient and fairs poorly. Under the DRG model, Hospital 11 has no comparable hospitals, even though its combination of inputs and outputs hardly changed. Adjusting for case mix altered the outputs for Hospital 11's peers and as a result, these hospitals were no longer comparable. A potential reason for this is that the case-mix adjustment factor for Hospital 11 is fairly large, in comparison to the case-mix adjustment factors of other small hospitals. Since there are no peer hospitals, Hospital 11 is classified as fully efficient. The case of Hospital 11 illustrates that a change in the make-up of comparable hospitals, when there is no change to the hospital itself, can greatly alter the evaluation of a hospital's efficiency.

As discussed at the start of this section, for hospitals with a case-mix adjustment factor of less than one, the efficiency score under the standard model is expected to be higher in comparison to the efficiency score under the DRG model. However, this is not always the case. The relationship between the case-mix adjustment factors and the efficiency scores under the standard model and the DRG model are explored in greater detail in Table 7.6. Table 7.6 lists individual hospitals' efficiency scores for each model, as well as the case-mix adjustment factor. Using this information, hospitals that behave unexpectedly are identified. It is important to note that efficiency scores and case-mix adjustment factors have been rounded to two decimal places in Table 7.6. As a result, it is possible that although the efficiency scores appear equal at two decimal places, they

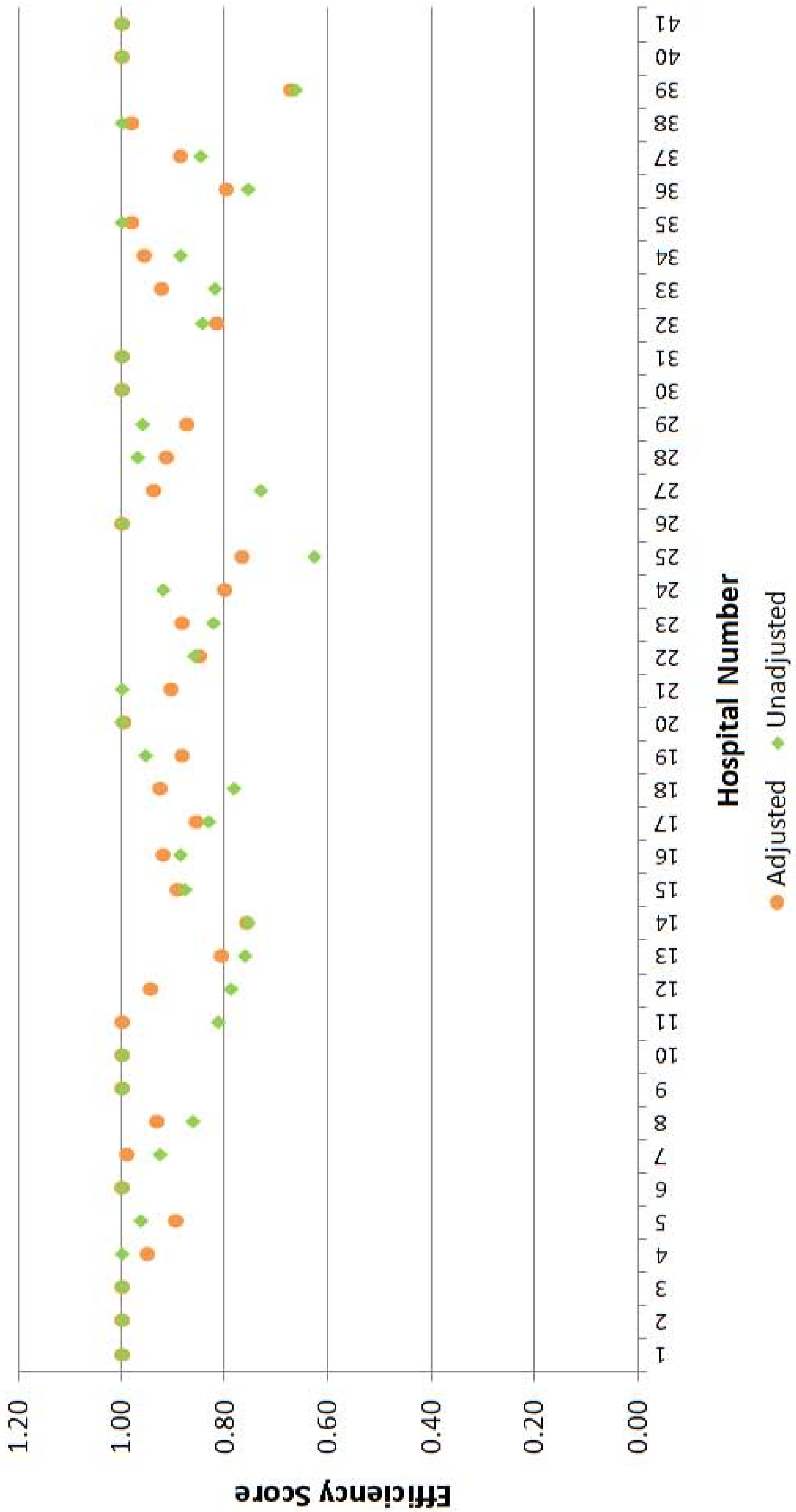

Figure 7.3: Efficiency scores for the DRG model and the standard model in 2011.

are not equal.

There are eight hospitals, shaded in gray in Table 7.6, that behave unexpectedly, in that, despite the case-mix adjustment factor being less than one, the adjusted efficiency score is greater than the unadjusted efficiency score. As explained for Hospital 11, above, the likely reason for this interaction is the change in relative performance. Hospital 8 was examined in more detail to understand this interaction.

Under the standard model, Hospital 8 has four peers, namely Hospital 2, Hospital 4, Hospital 6 and Hospital 9. That is, Hospital 8 is inefficient under the standard model relative to these four hospitals. The peers are almost identical under the DRG model, except that there are only three peers, namely Hospital 2, Hospital 6 and Hospital 9. The adjustment for case mix resulted in Hospital 4 no longer being comparable to Hospital 8. The case-mix adjustment factor for Hospital 8 is higher than those for Hospitals 2, 6 and 9. Intuitively, one would expect Hospital 8 to perform better relative to these hospitals after adjusting for case mix because it has a heavier case mix. However, this is not the case. The case-mix adjustment factor impacts the absolute number of cases, which in turn impacts the calculation of the relative use of inputs per unit of output. So although Hospital 8 has the heaviest case mix, it still uses relatively more inputs to produce a single output, in comparison to its peers. Furthermore, the use of inputs to produce a single output is higher under the DRG model than it is under the standard model. Like Hospital 11, Hospital 8 provides good example of how DEA evaluates efficiency relative to a hospital's peers.

Interestingly, there are no hospitals for which the case-mix adjustment factor is greater than one and the unadjusted score is greater than the adjusted efficiency score.

Hospital 25, the least efficient hospital under the standard model in 2011, experiences a large increase in efficiency, after adjusting for differences in case mix. Hospital 25 also has the largest case-mix adjustment factor. Clearly the efficiency score under the standard model was being confounded by a heavier case mix.

Four hospitals were identified as having efficiency scores under 0.8 in both the standard and DRG models, namely Hospital 25, Hospital 39, Hospital 36 and Hospital 14. For these four hospitals, even after taking differences in the mix of cases into account, large inefficiencies still exist.

It is clear from these results that efficiency scores will be biased if no adjustment is made for differences in case mix and it cannot be reasonably assumed that case mix is constant across the group of hospitals being analysed.

Having established a need for adjusting for differences in case mix, the results of different techniques of adjusting for case mix are analysed in the following section.

Table 7.6: Identification of hospitals that behave unusually after adjusting for case-mix.

Hospital	Unadjusted	Adjusted	DRG CMAF	Classification	Comparison
4	1.00	0.95	0.59	CMAF <1	Unadjusted Greater
3	1.00	1.00	0.59	CMAF <1	Equal
6	1.00	1.00	0.59	CMAF <1	Equal
5	0.96	0.90	0.67	CMAF <1	Unadjusted Greater
2	1.00	1.00	0.69	CMAF <1	Equal
9	1.00	1.00	0.69	CMAF <1	Equal
8	0.86	0.93	0.76	CMAF <1	Adjusted Greater
7	0.93	0.99	0.79	CMAF <1	Adjusted Greater
1	1.00	1.00	0.84	CMAF <1	Equal
24	0.92	0.80	0.84	CMAF <1	Unadjusted Greater
14	0.76	0.76	0.88	CMAF <1	Adjusted Greater
19	0.95	0.88	0.89	CMAF <1	Unadjusted Greater
21	1.00	0.91	0.89	CMAF <1	Unadjusted Greater
32	0.84	0.82	0.89	CMAF <1	Unadjusted Greater
10	1.00	1.00	0.90	CMAF <1	Equal
29	0.96	0.88	0.91	CMAF <1	Unadjusted Greater
35	1.00	0.98	0.93	CMAF <1	Unadjusted Greater
22	0.86	0.85	0.93	CMAF <1	Unadjusted Greater
38	1.00	0.98	0.94	CMAF <1	Unadjusted Greater
13	0.76	0.81	0.94	CMAF <1	Adjusted Greater
15	0.88	0.89	0.95	CMAF <1	Adjusted Greater
28	0.97	0.92	0.95	CMAF <1	Unadjusted Greater
20	1.00	1.00	0.95	CMAF <1	Unadjusted Greater
17	0.83	0.85	0.98	CMAF <1	Adjusted Greater
16	0.89	0.92	0.98	CMAF <1	Adjusted Greater
30	1.00	1.00	0.99	CMAF <1	Equal
11	0.81	1.00	1.00	CMAF <1	Adjusted Greater
39	0.66	0.67	1.02	CMAF >1	Adjusted Greater
26	1.00	1.00	1.02	CMAF >1	Equal
34	0.89	0.96	1.03	CMAF >1	Adjusted Greater
23	0.82	0.88	1.03	CMAF >1	Adjusted Greater
31	1.00	1.00	1.03	CMAF >1	Equal
37	0.85	0.89	1.05	CMAF >1	Adjusted Greater
36	0.75	0.80	1.06	CMAF >1	Adjusted Greater
12	0.79	0.95	1.09	CMAF >1	Adjusted Greater
18	0.78	0.93	1.11	CMAF >1	Adjusted Greater
41	1.00	1.00	1.12	CMAF >1	Equal
33	0.82	0.92	1.16	CMAF >1	Adjusted Greater
40	1.00	1.00	1.21	CMAF >1	Equal
27	0.73	0.94	1.25	CMAF >1	Adjusted Greater
25	0.63	0.77	1.43	CMAF >1	Adjusted Greater

7.4 A Comparison of the Different Case Mix Adjustment Techniques

7.4.1 Including Case-Mix Adjustment Factors as an Additional Output

For the additional output model, rather than adjust admissions using the case-mix adjustment factors, the factors were used as an additional output variable. The efficiency scores for the additional output model are expected to be higher than those of the DRG model because the number of variables used and resulting efficiency scores are positively correlated (Nguyen and Coelli, 2009), as explained in Section 4.2. As a result, the results of the DRG model and the additional output model are not directly comparable. So although the efficiency scores for the additional output model are higher, this does not necessarily mean that there is a higher absolute level of efficiency.

A summary of the results of the additional output and DRG model can be found in Table 7.7.

Table 7.7: Summary of efficiency scores for the additional output and the DRG model

	Additional Output Model		DRG Model	
Year	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient
2008	0.95	46.34%	0.91	31.71%
2009	0.96	53.66%	0.93	39.02%
2010	0.96	53.66%	0.93	34.15%
2011	0.96	53.66%	0.92	29.27%

As expected, average efficiency scores are higher for the additional output model. The proportion of fully efficient hospitals is also significantly higher. In 2011, over 50% of hospitals are fully efficient, whilst only about 30% are fully efficient under the DRG model.

Figure 7.4 illustrates the 2011 efficiency scores for both the DRG model and the additional output model. Although the efficiency scores are not directly comparable, Figure 7.4 is a good illustration of the pushing up and flattening of efficiency scores, as a result of the additional output variable. By adding an output to the model, the heterogeneity of each hospital is increased in that the combinations of inputs available to produce a combination of outputs increases. Therefore it is harder to find comparable hospitals and more hospitals are likely to be classified as fully efficient because there are no peers.

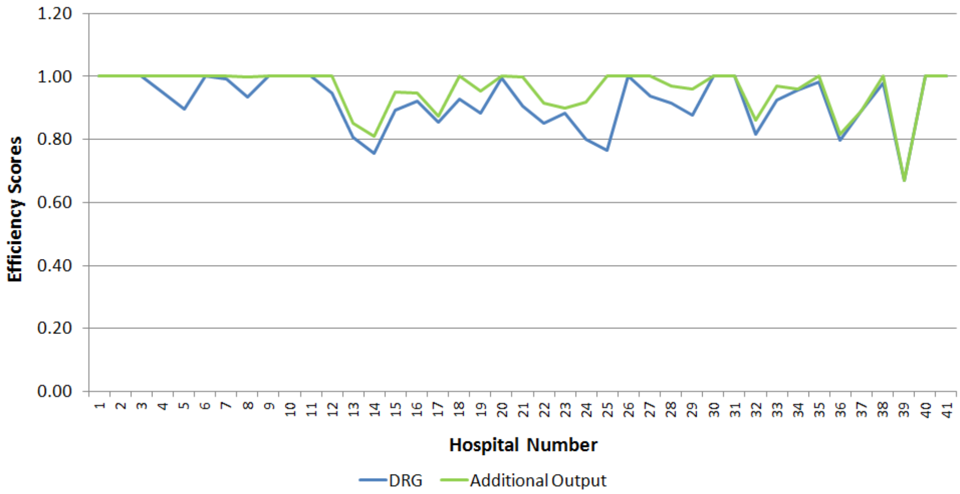

Figure 7.4: A comparison of efficiency scores for the DRG model and the additional output model in 2011 .

The very small hospitals, ranging from Hospital 1 to Hospital 12, all display the increasing and flattening of efficiency scores. Out of these 12 hospitals, only Hospital 8 has peers. Therefore, the other 11 have been classified as efficient because there are no comparable hospitals, which is as a result of the increased heterogeneity explained above. This is also the case for Hospital 25 which was one of the least efficient hospitals under the DRG model, however is fully efficient under the additional output model.

On the other side of the scale, the efficiency scores of the very large hospitals, ranging from Hospital 35 to Hospital 41, are almost identical for the two techniques. For these hospitals, the introduction of the additional input did not impact the comparability to the same extent as it influenced the small hospitals.

The three most inefficient hospitals under the DRG model, excluding Hospital 25, are still amongst the least efficient hospitals under the additional output model, namely Hospital 39, Hospital 14 and Hospital 36. The efficiency scores for these three hospitals are higher under the additional output model, although only marginally so for Hospital 39.

Efficiency scores for all hospitals over the four year period of the analysis for the additional output model can be found in Appendix A, Table A.3.

Only two hospitals are identified as efficient and have peer hospitals in their ERS's under the additional output model, namely Hospital 18 and Hospital 27. Under the DRG model, these hospitals had efficiency scores of around 0.93 and 0.94, respectively. The ERS for the additional output model for 2011 is shown in Figure 7.5.

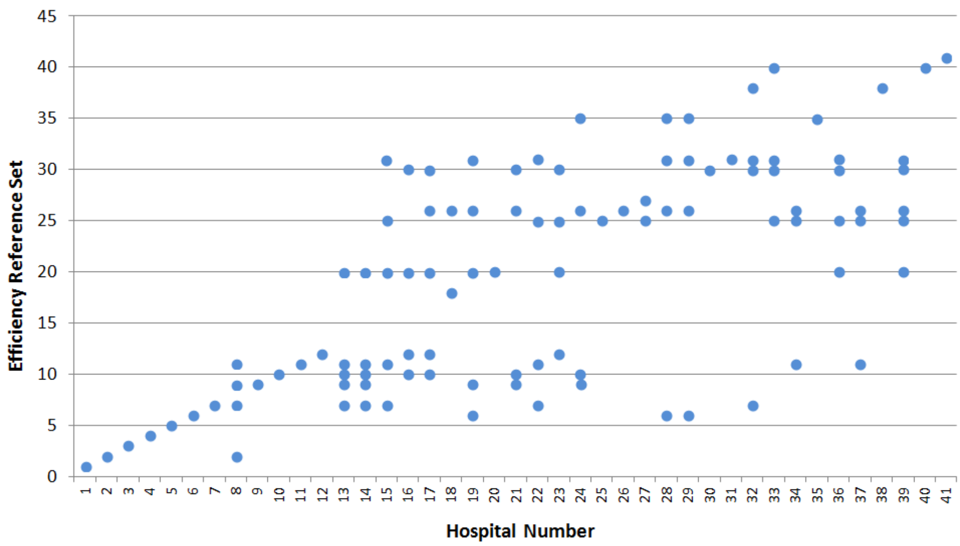

Figure 7.5: Illustration of the ERS for the additional output model.

It is clear from Figure 7.5 that the additional variable has increased the heterogeneity of hospitals and the ability to locate comparable hospitals has decreased. This is particularly clear for the very small hospitals, Hospital 1 to Hospital 12, and the two very large hospitals, Hospital 40 and Hospital 41, where they each make up their own ERS and this creates a diagonal line.

In summary, from the results it can be seen that the additional output distorts efficiency scores as a result of increased heterogeneity and inability to find comparable hospitals. That said, the hospitals that were most inefficient under the DRG model are still identified as inefficient under the additional output model, just not to the same extent. The exception to this is Hospital 25 which is one of the least efficient hospitals under the DRG model but fully efficient under the additional output model. This highlights the danger of using an additional output when a single case-mix adjusted output could be used instead.

7.4.2 Disaggregated Admissions

As suggested by Sherman (1984) and Grosskopft and Valdmanis (1993), admissions were disaggregated into broad categories to account for differences in case mix across hospitals. As mentioned in Section 7.4.1 above, the efficiency scores from the DRG model are not directly comparable to the efficiency scores of the disaggregated model, because of the different number of output variables. The efficiency scores of the disaggregated model are expected to be significantly higher than those of the DRG model. Furthermore, there is an expectation that there will be a flattening out of the efficiency scores at one, as the heterogeneity of outputs has increased and it is harder to find comparable hospitals. This is clearly visible in Figure 7.6.

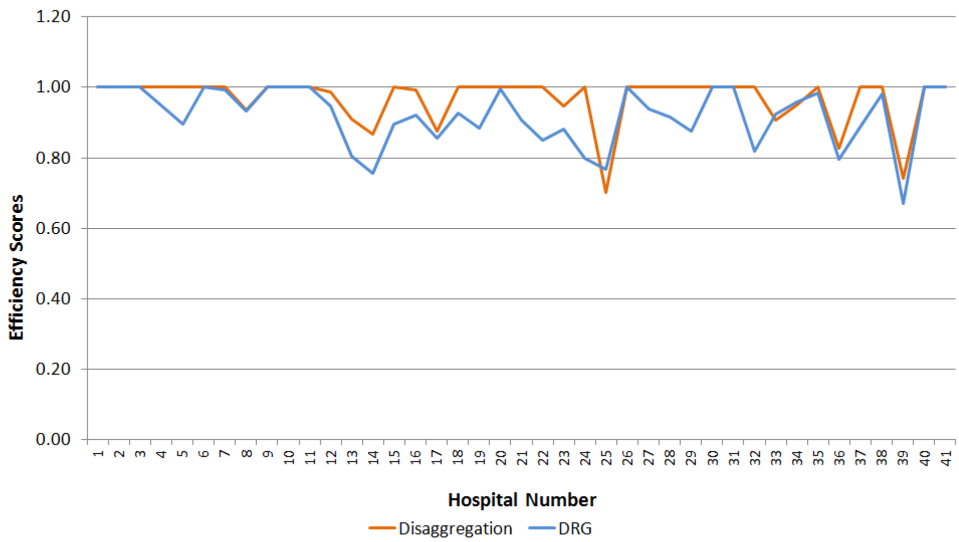

Figure 7.6: Efficiency scores for the DRG model and the disaggregated model for 2011.

The flattening of efficiency scores is most visible for Hospital 4 and Hospital 5, as well as Hospital 18 to Hospital 22 and Hospital 26 to Hospital 32. Interestingly, the disaggregated admissions efficiency score is not higher than the DRG efficiency scores for every hospital, despite the large difference in the number of output variables, for example Hospital 25 and Hospital 33. This suggests that the DRG case-mix adjustment factor may be capturing differences in case mix that are not identified through the process of disaggregation.

Consider the five hospitals with efficiency scores lower than 0.8 in the DRG model, namely Hospital 39, Hospital 14, Hospital 25, Hospital 36 and Hospital 24, all these hospitals are still identified as relatively inefficient under the disaggregated model, in spite of the increase in the number of outputs.

As under the additional output model, there are a number of hospitals that are relatively inefficient under the DRG model that are identified as fully efficient under the disaggregated model, for example Hospital 22 and Hospital 32. This is as a result of the increased heterogeneity and therefore lack of comparable hospitals.

A summary of the efficiency scores of the disaggregated model and the DRG model is contained in Table 7.8. The flattening out of efficiency scores at one is increasingly obvious when analysing the average efficiency scores and the proportion of hospitals which are fully efficient for the disaggregated model. The average efficiency scores are significantly higher for the disaggregated model, in comparison to the DRG model, and the proportion of hospitals which are fully efficient are roughly 70%, more than double the proportion under the DRG model. Efficiency scores for all the hospitals over the four year period for the disaggregated model can be found in Table A.4, in Appendix A.

Table 7.8: Summary of efficiency scores for the disaggregated output and the DRG model

	Disaggregated Output Model		DRG Model	
Year	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient
2008	0.97	70.73%	0.91	31.71%
2009	0.97	68.29%	0.93	39.02%
2010	0.97	70.73%	0.93	34.15%
2011	0.97	70.73%	0.92	29.27%

From the analysis of individual results it can be seen that, unlike the DRG model, two hospitals are truly efficient, in that they are identified as fully efficient relative to other hospitals, under the disaggregated model. These hospitals are Hospital 20 and Hospital 32.

Figure 7.7 illustrates the ERS's for each of the hospitals. The increased heterogeneity of outputs and, as a result, the decreased ability to find comparable hospitals is very clear, with a great many hospitals having no peers, as seen by the diagonal arrangement of points.

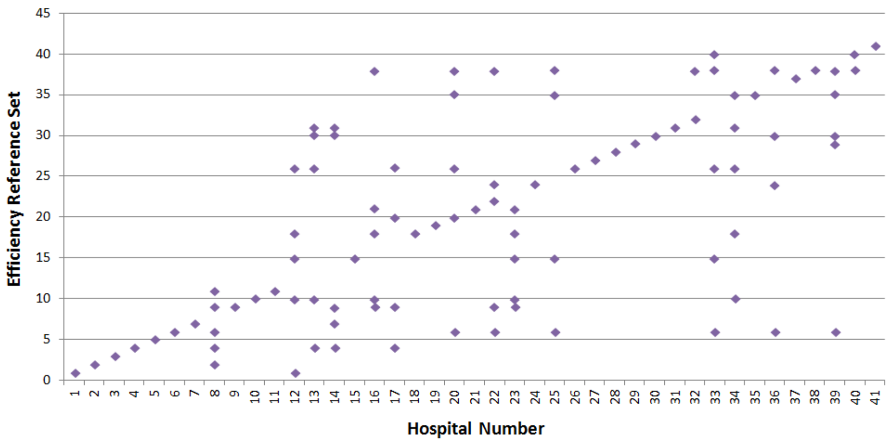

Figure 7.7: An illustration of the ERS's for the disaggregated-admissions model for each hospital in 2011.

There is not much evidence to suggest that the disaggregation of admissions captures case mix any better or any worse than the DRG model. However, the additional variables do make it harder to identify comparable hospitals which biases the efficiency scores upwards and makes it difficult to identify inefficient hospitals.

7.5 A Comparison of the DRG Model and the BDRG Model

Results of these two models are directly comparable, as they have the same number of input and output variables and the outputs have been adjusted for differences in case mix. Since very little difference was noted between the DRG adjustment factors and the BDRG adjustment factors, very little difference is expected between the results of each model.

The average efficiency scores and the proportion of hospitals that are fully efficient, as seen in Table 7.9, is almost identical to the summary results of the DRG model. Efficiency scores for all hospitals over the four year period of the analysis for the BDRG model can be found in Appendix A, Table A.5.

Table 7.9: Summary of efficiency scores for the BDRG and the DRG model

	BDRG Model		DRG Model	
Year	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient	Average Efficiency Score	Proportion of Hospitals which are Fully Efficient
2008	0.92	39.02%	0.91	31.71%
2009	0.93	36.59%	0.93	39.02%
2010	0.93	34.15%	0.93	34.15%
2011	0.93	34.15%	0.92	29.27%

Average efficiency scores are very close and both models display a marginal increase over the four year period. Roughly a third of hospitals across each of the four years were identified as being fully efficient in both of the models.

A comparison of the efficiency scores derived for each hospital for the DRG and BDRG models is shown in Figure 7.8.

For the majority of the hospitals, there appears to be very little difference between the efficiency scores produced by the BDRG model and the DRG model for 2011. However, small differences are present for some hospitals, for example Hospital 13 and Hospital 39. This may be a result of one case-mix adjustment factor detecting differences that are not distinguished by the other case-mix adjustment factor. In other words, for a

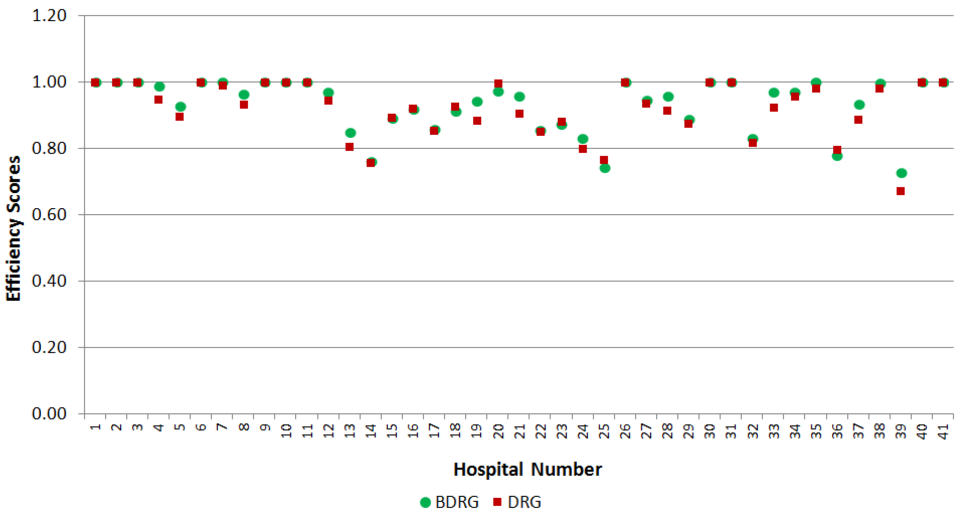

Figure 7.8: A comparison of efficiency scores for the BDRG model and the DRG model in 2011.

particular case, the BDRG case-mix adjustment factor may provide a better indication of the expected resource utilisation, in comparison to the DRG case-mix adjustment factor. For hospitals that have a high proportion of those cases, the BDRG case-mix adjustment factor for that hospital will be higher than the DRG case-mix adjustment factor. As a result, the efficiency score from the BDRG model is likely to be higher, since the output is higher, for the same level of inputs, however this does depend on a hospitals peers. The same concept is true for hospitals with a large proportion of cases for which the DRG case-mix adjustment factor is a better indicator of resource consumption and, therefore, the hospital DRG case-mix adjustment factor is larger than the BDRG case-mix adjustment factor. As a result, it is almost impossible to determine which adjustment factor captures differences in case mix better.

Table 7.10: Five hospitals with the lowest efficiency scores for the BDRG model in 2011.

Hospital Number	2008	2011
39	0.76	0.73
25	0.68	0.74
14	0.91	0.76
36	0.80	0.78
32	0.83	0.83

The hospitals identified as most efficient and as least efficient appeared to be approximately the same across the two models. Table 7.10 list the five hospitals that are identified as having the lowest efficiency scores in 2011 for the BDRG model. These five hospitals are identical to those under the DRG model, except for Hospital 32 whose rank was marginally higher under the DRG model. Hospital 39 is the worst performing hospital in both the BDRG model and the DRG model.

Unlike the DRG model, three hospitals were identified as fully efficient relative to peers under the BDRG model. Under the DRG model, no hospitals were identified as fully efficient relative to peers. The three hospitals are Hospital 9, Hospital 31 and Hospital 35. Hospital 9 and Hospital 31 are both efficient under the DRG model but have no peers, whilst Hospital 35 is marginally inefficient under the DRG model.

The ERS's for the two models for 2011 are illustrated in Figure 7.9. There are very few hospitals for which the ERS's are different.

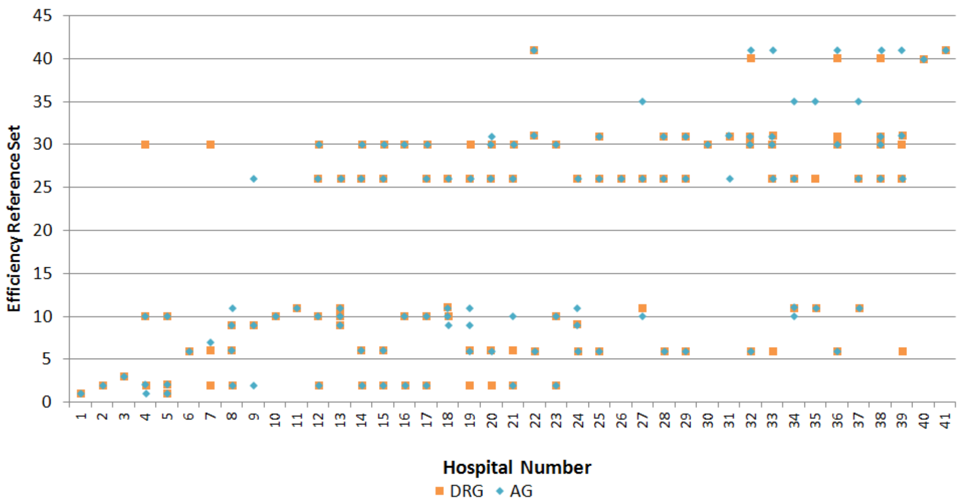

Figure 7.9: A comparison ERS's for the two case-mix-adjusted-admissions models.

It is clear that the refinement of the case-mix adjustment factor does not have a major impact on the efficiency scores. However, it is evident that the two case-mix adjustment factors are measuring different aspects of resource utilisation. The DRG measurement focuses on severity of illness and complications within a broad diagnosis category, whilst BDRG measurement focuses on age and gender within the same broad diagnosis category. As a result of focusing on different aspects, different hospitals will be favoured depending on the choice of the case-mix adjustment factor. However, this effect is minimal, as seen in Figure 7.8. A useful extension to this research would be to incorporate severity, age and gender all together in a case-mix adjustment factor. Unfortunately, there was insufficient data to do so for this analysis.

7.6 A Study of Individual Hospitals

Individual hospitals were chosen to be investigated further where there were interesting or unexpected results. Hospitals were selected by analysing key measures for 2011, namely the unadjusted efficiency scores, the DRG adjusted efficiency score, the DRG case-mix adjustment factor, the change in pure technical efficiency from 2008 to 2011, and the existence of peers. These key measures can be found in Table B.1, in Appendix B.

Four hospitals were chosen to be analysed individually. These hospitals, along with the reason for selecting each one, are listed in Table 7.11.

Table 7.11: Hospitals selected for individual analysis.

Hospital	Reason
Hospital 25	Lowest unadjusted efficiency score, one of lowest DRG adjusted efficiency scores and highest case-mix adjustment factor.
Hospital 39	Lowest DRG adjusted efficiency score and second lowest unadjusted efficiency score.
Hospital 34	Largest increase in pure technical efficiency from 2008 to 2011 based on DRG efficiency scores.
Hospital 14	Largest decrease in pure technical efficiency from 2008 to 2011 and second lowest DRG adjusted efficiency score in 2011.

Hospital 25

Hospital 25 had the lowest unadjusted inefficiency score in 2011 (0.63), a low DRG adjusted efficiency score (0.77) and the highest DRG case-mix adjustment factor (1.43). Furthermore, Hospital 25 was identified in the data analysis as problematic because of the very low number of admissions, in comparison to hospitals of similar size. There was also a decrease of over 10% in the number of operational beds from 2008 to 2011.

Table 7.12 lists ratios for Hospital 25, as well as its peer hospitals, as identified in Figure 7.2.

Hospital 25 appears to perform as well as its peer hospitals for all but one of these statistics, namely the average length of stay (LOS). The average LOS for Hospital 25 is more than double that of its peer hospitals. It was mentioned previously that Hospital 25 has a private HIV clinic, therefore it is expected to have a large proportion of patients that are HIV positive, who therefore require longer stays. However, further investigation of the clinical data is required to confirm this. It is widely accepted that HIV positive patients, who have compromised immune systems, often take far longer to recover than an HIV negative patient. This is a possible explanation for the significantly longer average

Table 7.12: Hospital 25 compared to peer hospitals in 2011.

Hospital	Occupancy	Average LOS	Operational Beds	Nurses per Bed	Pharmacy Index	DRG Case-Mix Adjustment Factor
Hospital 25	73.39%	6.15	166	1.10	0.91	1.43
Hospital 6	59.43%	2.41	30	0.77	0.47	0.59
Hospital 26	75.17%	2.76	178	1.36	0.94	1.02
Hospital 31	79.27%	3.63	192	1.14	0.78	1.03

LOS.

Even after adjusting for differences in case mix using a case-mix adjustment factor to adjust admissions, Hospital 25 still appears relatively inefficient with a score of 0.77 in 2011 under the DRG model. This may indicate that the case-mix adjustment factor is not capturing the true resource consumption of this mix of cases. The DRGs were inspected and it was found that a patient is classified under the opportunistic infection and there is no indicator for HIV. As a result, there is no indicator to account for differences in resource consumption due to HIV. Therefore, the resource consumption for these patients may be understated by the DRG case-mix adjustment factor. For example, when calculating the factor for a particular group, the HIV cases will be significantly higher than the average billed amount and this average will not be a good indication of the resource consumption of one of these cases. The overall hospital case-mix adjustment factor will be understated if there are a large proportion of cases in these individually understated groups.

Hospital 25 may not appear quite so inefficient if case-mix adjustments incorporating HIV could be made.

Hospital 39

Hospital 39 had the lowest efficiency score under the DRG model (0.67) and the second lowest efficiency score under the standard model (0.66). There are two noticeable contributing factors to this relatively low efficiency. Firstly, Hospital 39 has one of the highest pharmacy ratios for 2011, utilising almost 20% more than what would have been expected. Secondly, the number of admissions in 2011 was substantially lower than hospitals of similar size.

Hospital 39 is a large hospital and grew by almost 15% over the four year period of the analysis, when hospital size is measured by number of operational beds, this growth is a contributing factor to the large decrease in efficiency from 2008 to 2011, as seen in Table 7.3.

Table 7.13 lists ratios for Hospital 25, as well as its peer hospitals, as identified in Figure 7.2.

Table 7.13: Hospital 39 compared to peer hospitals in 2011.

Hospital	Occupancy	Average LOS	Operational Beds	Nurses per Bed	Pharmacy Index	DRG Case-Mix Adjustment Factor
Hospital 39	73.39%	3.77	285	1.25	1.21	1.02
Hospital 6	59.43%	2.41	30	0.77	0.47	0.59
Hospital 26	75.17%	2.76	178	1.36	0.94	1.02
Hospital 30	85.14%	2.92	191	1.74	0.93	0.99
Hospital 31	79.27%	3.63	192	1.14	0.78	1.03

Comparing these ratios, it appears that the pharmacy index is one of the primary reasons for Hospital 39 appearing inefficient relative to its peers. Whilst the average LOS is similar to that of Hospital 31, Hospital 31 balances this by using relatively fewer nurses per bed and a much lower use of pharmaceuticals in comparison to Hospital 39.

Stricter controls around use and prescription of pharmaceuticals would be a useful first step towards improving the efficiency of this hospital.

Hospital 34

Hospital 34 had the largest increase in pure technical efficiency between 2008 and 2011, increasing by 17.07%. In spite of the 17% increase, Hospital 34 was not fully efficient under the standard model or the DRG model. The inputs and outputs for Hospital 34 are shown for each year of the study in Table 7.14.

The case-mix adjusted number of admissions has remained relatively stable around 16 500 cases for each of the years. Whilst the output remained relatively constant, both the number of operational beds and the pharmacy index increased and the adjusted number of nurses decreased. In order for there to have been an efficiency improvement, the decrease in the adjusted number of nurses must have offset the relative increase in beds and pharmaceuticals consumed.

Since the adjusted number of cases have not varied by much over the years, the increase in the use of pharmaceutical goods, relative to what was expected, would seem

Table 7.14: Inputs and outputs of Hospital 34 for 2008 to 2011.

	Inputs			Outputs
Year	Beds	Pharmacy Index	Nurses	Adjusted Admissions
2008	194.00	1.13	346.39	16230.47
2009	194.00	1.10	255.88	16653.83
2010	212.63	1.15	223.81	16304.24
2011	208.00	1.16	237.26	16591.90

to be unexpected and unnecessary. This is an area in which input savings could be made in order to ensure that this hospital becomes more efficient.

The combination of increasing the number of operational beds and decreasing the adjusted number of nurses means that the number of nurses per bed has fallen over this period from 1.79 in 2008 to 1.14 in 2011. Whilst the ratio of 1.14 is not out of line with those ratios of efficient hospitals, the impact that this decrease has had on the quality of care should be investigated.

Hospital 14

Hospital 14 had the largest decrease in pure technical efficiency between 2008 and 2011, decreasing by 15.60%. In 2008, Hospital 14 had an efficiency score of 0.91 under the DRG model, which decreased to 0.76 in 2011. The efficiency scores appeared to remain relatively stable in 2009 and 2010 with the decrease occurring in 2011. In order to understand this sudden decrease, the inputs and outputs for Hospital 34 for each of the years of the study are examined. Theses inputs and outputs are shown for each year of the study in Table 7.15.

Table 7.15: Inputs and outputs of Hospital 14 for 2008 to 2011.

	Inputs			Outputs
Year	Beds	Pharmacy Index	Nurses	Adjusted Admissions
2008	63.50	0.99	112.25	6180.15
2009	70.67	0.95	114.05	6537.14
2010	72.00	0.90	116.92	6386.17
2011	99.00	0.95	125.77	6958.71

There is no obvious reason driving the decrease in technical efficiency from 2010 to 2011. All of the inputs increased, as did the outputs. The percentage change for inputs and outputs from 2010 to 2011 was calculated. The number of operational beds increased by 37.50% from 2010 to 2011, the largest increase in inputs and outputs over this period.

The pharmacy index went up by 5.56%, the number of nurses increased by 7.57%, whilst outputs increased by 8.97%. Looking at these increases only, it becomes evident that the increased number of beds is driving the decrease in efficiency from 2010 to 2011. This is confirmed by the decrease in occupancy from 79.28% in 2008 to 59.12% in 2011.

When compared to its peers in 2011, it is less clear that occupancy is driving the relative inefficiency; rather, it appears to be a combination of factors. Ratios for Hospital 14 and its peers are listed in Table 7.16.

Table 7.16: Hospital 14 compared to peer hospitals in 2011.

Hospital	Occupancy	Average LOS	Operational Beds	Nurses per Bed	Pharmacy Index	DRG Case-Mix Adjustment Factor
Hospital 14	59.12%	2.63	99	1.27	0.95	0.88
Hospital 2	77.21%	2.48	21	0.83	0.66	0.69
Hospital 6	59.43%	2.41	30	0.77	0.47	0.59
Hospital 26	75.17%	2.76	178	1.36	0.94	1.02
Hospital 30	85.14%	2.92	191	1.74	0.93	0.99

The occupancy rate for Hospital 14 is almost identical to the occupancy rate for Hospital 6. However, Hospital 14 has significantly more nurses per operational bed in comparison to Hospital 6, and uses significantly more pharmaceuticals. A similar ratio of pharmaceutical goods is used in comparison to Hospital 26 and Hospital 30, however the occupancy rate for Hospital 14 is significantly lower in comparison to these two hospitals.

The driving factors of inefficiency are the combination of a low occupancy rate coupled with a relatively high number of nurses per bed. A useful first step towards improving the efficiency of Hospital 14 would be to investigate the reason for low occupancy. If occupancy remains low, the necessity of a relatively large number of nurses per bed should be considered. Whilst the pharmaceutical use is lower than expected, the very low pharmacy index for the peers suggest that further reductions are possible.

Chapter 8

Conclusions and Further Research

8.1 Conclusion

This dissertation investigates hospital efficiency, with the primary aim of analysing the impact that adjusting for differences in case mix has on efficiency scores. There are no existing South African studies on hospital efficiency which incorporate an adjustment for case mix. This was one of the primary motivations for this dissertation and it is hoped that this study will contribute to this limited field of research in South Africa.

DEA was chosen as the efficiency measurement method because of its flexibility and ease of handling multiple inputs and outputs, which was particularly useful when investigating methods of incorporating a case-mix adjustment. Furthermore, no *a priori* assumptions regarding the distribution of the production frontier and the inefficiency term are needed when using DEA.

A number of DEA models were applied to a sample of South African private hospitals for the years 2008 to 2011. The data relate to only one of the three major hospital groups in South Africa; consequently, results are not necessarily indicative of the efficiency of the private hospital industry as a whole. Rather, they relate to this hospital group in particular. However, the three private hospital groups are structurally very similar therefore, the methodology used could easily be applied to hospitals in the private sector. The methodology could also be extended to the public sector, provided adequate data are available and methodological decisions are reviewed.

Five DEA models, using three identical inputs and different combinations of outputs were used to evaluate the impact that an adjustment for case mix has on efficiency scores. The results of the DEA models were used to identify hospitals that are fully efficient and the extent to which resource savings can be made, as well as to identify hospitals that behaved unexpectedly or unusually.

A model with no adjustment for differences in case mix across hospitals, the standard model, was compared to a Diagnosis-Related Group (DRG) case-mix adjusted model, the

DRG model, to determine the effect of including a case-mix adjustment in the assessment of hospital efficiency. The DRG model used case-mix adjusted admissions, calculated using a DRG case-mix adjustment factor. The DRG model is consistent with the international literature and is thought to be most representative of the hospital production process. For this reason, the DRG model and the comparison to the standard model was the primary focus of the investigation. Little difference was noted between the average efficiency scores produced by the DRG and standard model in 2011; the average efficiency score in 2011 was 92% under the DRG model and 90% under the standard model. This is only a 2% difference in average resource savings. However, whilst little difference was noted in the average efficiency scores, there are substantial differences between the adjusted and unadjusted efficiency scores of individual hospitals, as seen in Figure 7.3. It is clear that omitting an adjustment for case mix biases efficiency scores and under the standard model the efficiency scores for some hospitals are overstated, whilst for others they are understated.

It is important to note that the interaction between case mix and efficiency scores is not as simple as concluding that a heavier case mix equates to a higher adjusted efficiency score, in comparison to an unadjusted efficiency score, as demonstrated by the results attained. When using a DEA model, hospitals are evaluated relative to one another, therefore a change in either the hospital being evaluated or a change to one of the comparable hospitals can result in a change in the relative efficiency. This is one of the advantages of DEA and one of the reasons why DEA models are used in complex production environments, as this interaction would be very difficult to assess when using ratio analysis alone. However, this does mean that efficiency scores are not stable. In other words, relatively small changes to one hospital can have a significant impact on the efficiency scores of multiple hospitals.

Three different methods of adjusting for case mix were explored, namely a case-mix adjustment factor used to adjust outputs, a case-mix adjustment factor as an additional variable and finally disaggregation of admissions. The loss of power of the DEA model as a result of increasing the number of outputs is apparent when looking at the average efficiency scores of the three models. Under the DRG model when there is only one output, the average efficiency score in 2011 is 92%. This jumps to 96% when the case-mix adjustment factor is used as a second output. When admissions are disaggregated into five separate outputs, the average efficiency score increases to 97%. Whilst the choice of inputs and outputs should be fully representative of the production process, this limitation of DEA should be kept in mind and multiple outputs should be balanced by using a large number of hospitals in the sample being analysed.

If there is sufficient data to construct a case-mix adjustment factor, there is little sense including it as an additional variable which results in a decrease of power of the model. Case-mix adjusted admissions should be used as a single output instead. However, in the event that there is insufficient data to construct a case-mix adjustment factor, disaggregating admissions does appear to capture some differences in case mix. The results of the disaggregated admissions model show that many hospitals are classified as fully efficient as there are no comparable hospitals, around 70% of hospitals were classified as fully efficient in 2011. Therefore, the results of this model should be interpreted with caution.

Although every attempt has been made to adjust for differences in case mix, the risk still exists that the calculated case-mix adjustment factor does not adequately capture these differences. As a result, a refinement of the DRG case-mix adjustment factor was considered. The DRG classification uses diagnosis and severity of illness to divide cases into groups. In the refinement, age and gender were used as criterion, instead of severity of illness, and outliers were removed. There were minor differences noted at hospital level for these two case-mix adjustment factors and little difference in the resulting efficiency scores. It is important to note that for the purposes of this investigation, it was not necessary for case-mix adjustment factors to be accurate at diagnosis group level, only at hospital level. In applications where it is essential for group-level factors to be accurate, such as alternative reimbursement, refinements may make a significant difference.

Whilst the efficiency scores provide an indication of the resource savings possible within this group of hospitals, nothing can be concluded about the absolute efficiency of the group of hospitals. This is one of the major limitations of all efficiency measurement models.

A more efficient hospital industry is necessary in order to progress with health care reform in South Africa, as noted in the Green Paper on NHI by the Department of Health. Hospitals which are more efficient should be able to treat a specific number of patients using fewer inputs, which should lead to financial savings and could effectively assist in curbing increasing hospital costs. This could therefore have major implications, not only for this particular hospital group, but if the methodology is applied to the health care environment as a whole, for future health care costs.

It should be noted that in order to apply this methodology to include other hospitals, both in the public and private sector, extensive further research would be required. These areas of research are discussed briefly in the following section.

8.2 Further Research

Suggestions for further research primarily include two interesting extensions to this piece of research, which would provide a more complete picture of hospital efficiency in South Africa.

Firstly, a comparison of results for the three different hospital groups currently dominating the South African private hospital environment would provide useful insights into the private hospital industry as a whole, particularly from a benchmarking perspective. Furthermore, managerial best practices could be identified through recognising and analysing the most efficient hospitals, and applying these practices to other hospitals, with the overall aim of enhancing hospital efficiency in the private helath care sector, as a whole.

Another interesting extension to this research would be to apply this methodology to the public hospital industry. Such an efficiency analysis would provide a useful platform to analyse the extent of inefficiencies in the public hospital industry, and, moreover, to determine to what extent these inefficiencies differ from those observed in the private hospital industry.

In order to apply this methodology successfully to these areas of further research, the completion of three major tasks is essential. These tasks include the testing of methodological decisions and their appropriateness when including public sector hospitals in an analysis, incorporating a measure for quality of care and refining the case-mix adjustment factor.

Before this methodology could be applied to public sector hospitals in particular, a number of methodological decisions should be tested and refined. As indicated earlier, an input-orientation is appropriate in the context of private hospitals, as they are for-profit institutions. However, this assumption is inappropriate for public hospitals which are not profit driven. Alternative measurements of inputs and outputs should also be investigated, most notably a measurement for the flow of capital. In this piece of research, the number of beds was used as a proxy for capital, whilst not ideal, it was the best proxy available.

Quality of care was assumed to be constant across all the hospitals in the analysis, as all data was sourced from a single hospital group with the same quality standards across all hospitals. When incorporating a wider range of hospitals, it is essential to allow for quality of care in determining efficiency as relatively more resources may be used to provide better quality care. If no allowance is made for differences in quality of care, there is a risk that hospitals using relatively fewer resources to produce sub-standard care are classified as efficient. Therefore, a useful extension to this piece of research would be an investigation into the measurement of quality of care and the method of including it into an efficiency analysis.

A further area of research, in addition to those mentioned above, but perhaps equally important is that there needs to be a consideration of the case-mix adjustment factor, in order to ensure that it captures all the differences in case mix. Consequently, further research could be conducted on the refinement of a case-mix adjustment factor by including severity of illness in conjunction with age and gender as criterion for estimating resource consumption. Other factors, such as the presence of co-morbidities, should also be investigated. Such a refinement would require significantly more data than was available in this study.

In the final analysis, efficiency in health care in the South African context, which constantly faces the challenge of limited resources in the face of a burgeoning population, it is of critical importance. It cannot be denied that every available method of enhancing efficiency in such a critical area, namely the health care services offered to the South African population, should be examined with great diligence. This is particularly true, given the context of the impending introduction of the NHI, which will have implications for all South Africans on both an individual and a societal level.

Bibliography

Balk, B. M. (2001). Scale efficiency and productivity change, *Journal of Productivity Analysis* **15**: 159–183.

Banker, R. D., Charnes, A. and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Journal of Management Science* **30**(9).

Barnum, D. and Gleason, J. (2012). Technical efficiency bias in data envelopment analysis caused by intra-output aggregation, *Applied Economics Letters* **14**(9): 623–626.

Boulle, A., Blecher, M. and Burn, A. (2008). *Health and Related Indicators*, Health Systems Trust, Durban, pp. 231–250.

Caves, D. W., Christensen, L. R. and Diewert, W. E. (1982). Multilateral comparisons of output, inputs, and productivity using superlative index numbers, *The Economic Journal* **92**(365): 73–86.

Centre for Development and Enterprise (2011). Reforming Healthcare in South Africa: What role for the private sector?
URL: *http://www.cde.org.za/article.php?a_id=413*

Charnes, A., Cooper, W. W. and Rhodes, E. (1978). Measuring the efficiency of decision making units, *European Journal of Operational Research* **2**: 429–444.

Childs, B. (2009). Quantifying the potential capacity in the private hospital sector to absorb an influx of newly covered patients, *Private Hospital Review 2009*, chapter 3.

Chilingerian, J. A. (1994). Exploring Why Some Physisians' Hospital Practices Are More Efficient: Taking DEA Inside the Hospital, *Data Envelopment Analysis: Theory, Methodology, and Application*, Springer, New York.

Clement, J., Vladmanis, V., Bazzoli, G., Zhao, M. and Chukmaitov, A. (2008). Is more better? An analysis of hospital outcomes and efficiency with a DEA model of output congestion, *Health Care Management Science* **11**: 67–77.

Coelli, T. (1998). A multi-stage methodology for the solution of orientated DEA models, *Operations Research Letters* **23**(3-5): 143–149.

Coelli, T. J. (1996). A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program, *Centre for Efficiency and Productivity Analysis Working Paper Series* (8/96).

Coelli, T., Rao, D. P., O'Donnell, C. and Battese, G. (2005). *An Introduction to Efficiency and Productivity Analysis*, Springer, New York.

Competition Tribunal (2005). Medi-Clinic Investment (Pty) Ltd and Wits University Donald Gordon Medical Centre Pty (Ltd), (75/LM/Aug05).

Competition Tribunal (2006). Netcare Hospital Group Pty (Ltd) and Community Hospital Group Pty (Ltd), (68/LM/Aug06).

Competition Tribunal (2010). Life Healthcare Group Pry (Ltd) and Amabubesi Hospitals Pty (Ltd) and Bayview Private Hospital Pty (Ltd), (11/LM/Mar10).

Cook, W. D. and Zhu, J. (2007). Classifying inputs and outputs in data envelopment analysis, *European Journal of Operational Research* **180**: 692–699.

Costa, M. D., Buys, R. and Pyle, A. (2009). Cost benchmarking multidisciplinary private hospital costs, *Private Hospital Review 2009*, chapter 5.

Council for Medical Schemes (2008). *Research Brief 1 of 2008*, Council for Medical Schemes, Pretoria.

Council for Medical Schemes (2012). *Annual Report 2011 - 2012*, Council for Medical Schemes, Pretoria.

Debreu, G. (1951). The coefficient of resource utilisation, *Econometrica* **19**(3): 273–292.

Department of Health (1998). *Medical Schemes Act, Act. No. 131 of 1998*, Republic of South Africa, Pretoria.

Department of Health (2011a). *Green Paper on National Health Insurance in South Africa*, Republic of South Africa, Pretoria.

Department of Health (2011b). *National Health Act, 2003*, Republic of South Africa, Pretoria.

Econex (2010). Health reform note 9.
URL: *www.econex.co.za/images/stories/ECONEX_Health Reform Note_9.pdf*

Erlandsen, E. (2008). Improving the efficiency of health care spending: What can be learnt from partial and selected analyses of hospital performance, *OECD Economics Department Working Papers* (555).

Fare, R., Grosskopf, S., Lindgren, B. and Roos, P. (1989). Productivity development in swedish hospitals: A malmquist output index approach, *Data Envelopment Analysis: Theory, Methodology and Applications*, Kluwer Academic Publishers, pp. 253–272.

Färe, R., Grosskopf, S., Norris, M. and Zhang, Z. (1994). Productivity growth, technical progress, and efficiency changes in industrialised countries, *The American Economic Review* **84**(1): 66–83.

Farrell, M. J. (1957). The measurement of productive efficiency, *Journal of the Royal Statistical Society, Series A (General)* **120**(3): 253–290.

Fetter, R. B. (1991). Diagnosis related groups: Understanding hospital performance, *Interfaces* **21**(1): 6–26.

Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F. and Thompson, J. D. (1980). Case mix definition by diagnosis-related groups, *Medical Care* **18**(2, Supplement): 1–53.

Grosskopf, S. and Valdmanis, V. (1987). Measuring hospital performance: A non-parametric approach, *Journal of Health Economics* **6**: 89–107.

Grosskopft, S. and Valdmanis, V. (1993). Evaluating hospital performance with case-mix-adjusted outputs, *Medical Care* **31**(6): 525–532.

Gustafson, D. H. and Holloway, D. C. (1975). A decision theory approach to measuring severity in illness, *Health Services Research* **10**: 97–106.

Health Professions Council of South Africa (1974). *Health Professions Act, Act No. 56 of 1974*, Pretoria.

Heavens, J. (1999). *Casemix - The Missing Link in South African Healthcare Management*, South African Medical Research Council, Cape Town.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care, *Health Care Management Science* **6**: 203–218.

Hollingsworth, B., Dawson, P. and Maniadakis, N. (1999). Efficiency measurement of health care: a review of non-parametric methods and applications, *Health Care Management Science* **2**: 161–172.

Horn, S. D., Sharkey, P. D., Chambers, A. F. and Horn, R. A. (1985). Severity of Illness within DRGs: Impact on Prospective Payment, *American Journal of Public Health* **75**(10): 1195–1199.

Jacobs, R. (2001). Alternative methods to examine hospital efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis, *Health Care Management Science* **4**: 103–115.

Kao, C. and Liu, S.-T. (2009). Stochastic data envelopment analysis in measuring the efficiency of taiwan commercial banks, *European Journal of Operational Research* **196**(1): 312 – 322.

Kibambe, J. N. and Koch, S. F. (2007). DEA Applied to a Gauteng Sample of Public Hospitals, *South African Journal of Economics* **75**: 351–368.

Kinfu, Y. (2011). The efficiency of the health care system in South Africa: evidence from stochastic frontier analysis, *Applied Economics* .

Kirigia, J., Lambo, E. and Sambo, L. (2000). Are public hospitals in Kwazulu-Natal province of South Africa technically efficient?, *African Journal of Health Sciences* **7**: 67–77.

Kittelsen, S. A., Magnussen, J., Anthun, K. S., Häkkinen, U., Linna, M., Medin, E., Olsen, K. R. and Rhenberg, C. (2008). Hospital productivity and the Norwegian ownership reform - A Nordic comparative study, *University of Oslo, Health Economics Research Programm, Working Paper Series* (10).

Koopmans, T. C. (1951). An analysis of production as an efficient combination of activities, *Activity Analysis of Production and Allocation*, Wiley, New York.

Linna, M. (1998). Measuring Hospital Cost Efficiency with Panel Data Models, *Journal of Health Economics* **7**: 415–427.

Matsebula, T. and Willie, M. (2007). *Private Hospitals*, Health Systems Trust, Durban.

Mutter, R. L., Rosko, M. D. and Wong, H. S. (2008). Measuring hospital inefficiency: The effects of controlling for quality and paitent burden of illness, *Health Services Research* **43**: 1992–2013.

National Hospital Network (2012). About us.
URL: *http://www.nhn.co.za/ReadContent.aspx?id=8*

Nguyen, K.-H. and Coelli, T. (2009). Quantifying the effects of modelling choices on hospital efficiency measures: A meta-regression analysis, *Centre for Efficiency and Productivity Analysis Working Paper Series* (WP07/2009).

O'Neill, L., Rauner, M., Heidenberger, K. and Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies, *Socio-Economic Planning Sciences* **42**: 158–189.

Parkin, D. and Hollingsworth, B. (1997). Measuring productive efficiency of acute hospitals in Scotland, 1991 - 94: validity issues in Data Envelopment Analysis, *Applied Economics* **29**: 1425–1433.

Ramjee, S. and McLeod, H. (2010). *Private Sector Perspectives on National Health Insurance*, Health Systems Trust, Durban, pp. 179–194.

Rosko, M. and Carpenter, C. (1994). The impact of intra-drg severity of illness on hospital profitability: Implications for payment reform, *Journal of Health Politics, Policy and Law* **19**(4): 729–751.

Rosko, M. and Chilingerian, J. (1999). Estimating hospital efficiency: Does case mix matter?, *Journal of Medical Systems* **23**(1): 57–71.

Scheller-Kreinsen, D., Geissler, A. and Busse, R. (2009). The ABC of DRGs, *Euro Observer* **11**(4): 1–5.

Sherman, H. D. (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique, *Medical Care* **22**(10): 922–938.

Sherman, H. D. and Zhu, J. (2006). *Service Productivity Management: Improving Service Performance Using Data Envelopment Analysis (DEA)*, Springer, Boston.

Sloman, J. (2006). *Economics*, 6 edn, Financial Times Prentice Hall, London.

Söderland, N., Schierhout, G. and van der Heever, A. (1998). *Private health care in South Africa: Technical Report to Chapter 13 of the 1998 SA Health Review*, Health Systems Trust, Durban.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Vanderviere, E. and Huber, M. (2004). An adjusted boxplot for skewed distributions, *COMPSTAT'2004 Symposium* .

Wadee, H., Gilson, L., Thiede, M., Okorafor, O. and McIntyre, D. (2003). Health Care Inequality in South Africa and the Public/Private Mix, *United Nations Research Institute for Social Development* .

WHO (2009). South africa: health profile.
URL: *http://www.who.int/gho/countries/zaf.pdf*

Worthington, A. C. (2004). Frontier efficiency measurement in health care: A review of empirical techniques and selected applications, *Medical Care Research and Review* **61**(2): 135–170.

Zere, E., McIntyre, D. and Addison, T. (2001). Technical Efficiency and Productivity of Public Sector Hospitals in Three South African Provinces, *South African Journal of Economics* **69**: 336–358.

Zuckerman, S., Hadley, J. and Iezzoni, L. (1994). Measuring hospital efficiency with frontier cost functions, *Journal of Health Economics* **13**: 255–280.

Appendix A

Efficiency Scores

Table A.1: Efficiency scores for the DRG model for 2008 to 2011.

	2008	2009	2010	2011
1	1.00000	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000	1.00000
4	0.94374	0.96827	0.91650	0.94971
5	0.92790	0.94695	0.93048	0.89538
6	0.87158	0.96844	0.96792	1.00000
7	0.94199	0.99307	0.87698	0.99190
8	1.00000	1.00000	1.00000	0.93335
9	1.00000	0.93164	0.99682	1.00000
10	1.00000	1.00000	1.00000	1.00000
11	1.00000	0.96535	0.99957	1.00000
12	0.99005	1.00000	1.00000	0.94560
13	0.74828	0.83102	0.80374	0.80569
14	0.91012	0.90144	0.91073	0.75697
15	0.94803	1.00000	0.93357	0.89425
16	0.90827	0.93008	0.91638	0.92109
17	0.79861	0.83632	0.85515	0.85472
18	0.95180	0.87200	0.90836	0.92765
19	0.88574	0.95114	0.94251	0.88357
20	1.00000	1.00000	1.00000	0.99552
21	0.89096	0.94618	0.90427	0.90579
22	0.79404	0.78736	0.76174	0.85093
23	0.80309	0.92345	0.90739	0.88215
24	0.84637	0.83789	0.86634	0.79993
25	0.67968	0.72471	0.81735	0.76692
26	0.97447	1.00000	1.00000	1.00000
27	0.82938	0.80587	0.92038	0.93751
28	0.85921	0.92120	0.95099	0.91527
29	0.89186	0.93308	0.91778	0.87584
30	1.00000	1.00000	1.00000	1.00000
31	0.91541	0.86974	0.93712	1.00000
32	0.83345	0.84460	0.78848	0.81772
33	0.97772	1.00000	1.00000	0.92498
34	0.82028	0.93007	1.00000	0.95763
35	1.00000	1.00000	0.99847	0.98296
36	0.81383	0.85612	0.79568	0.79681
37	1.00000	1.00000	0.97250	0.88818
38	0.97780	1.00000	1.00000	0.98028
39	0.73023	0.74209	0.69288	0.67092
40	1.00000	1.00000	1.00000	1.00000
41	1.00000	1.00000	1.00000	1.00000

Table A.2: Efficiency scores for the standard model for 2008 to 2011.

	2008	2009	2010	2011
1	1.00000	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000	1.00000
4	0.97971	1.00000	1.00000	1.00000
5	0.94218	0.96822	0.96440	0.96189
6	0.92228	0.97896	0.97306	1.00000
7	0.94199	0.99307	0.87413	0.92593
8	1.00000	1.00000	0.98559	0.86072
9	1.00000	1.00000	1.00000	1.00000
10	0.97351	1.00000	1.00000	1.00000
11	0.93974	0.89330	0.83997	0.81359
12	0.88655	0.86824	0.87229	0.78826
13	0.77280	0.79744	0.76133	0.76015
14	0.94309	0.90524	0.88697	0.75523
15	0.91208	0.95043	0.86555	0.87821
16	0.93284	0.92239	0.89341	0.88804
17	0.80013	0.83463	0.85242	0.83073
18	0.78138	0.77815	0.78430	0.78349
19	0.93325	0.98882	0.97119	0.95282
20	1.00000	1.00000	1.00000	1.00000
21	0.93020	0.94756	0.95068	0.99920
22	0.82295	0.81349	0.77549	0.85989
23	0.85653	0.87435	0.85078	0.82249
24	0.95999	0.94998	0.94692	0.91913
25	0.65103	0.70720	0.78977	0.62791
26	0.97331	0.96946	1.00000	1.00000
27	0.70653	0.64870	0.65591	0.73006
28	0.86212	0.91170	0.90186	0.96886
29	0.99426	1.00000	0.99232	0.96075
30	0.99367	0.96316	1.00000	1.00000
31	0.91687	0.89031	0.87371	1.00000
32	0.86458	0.86896	0.84526	0.84468
33	0.94207	0.92012	0.85090	0.81976
34	0.77153	0.86835	0.90775	0.88781
35	1.00000	1.00000	1.00000	1.00000
36	0.80538	0.81759	0.75465	0.75402
37	0.94438	0.91702	0.87236	0.84670
38	1.00000	1.00000	1.00000	1.00000
39	0.72791	0.72891	0.67703	0.66251
40	1.00000	1.00000	1.00000	1.00000
41	1.00000	1.00000	0.96750	1.00000

Table A.3: Efficiency scores for the additional output model for 2008 to 2011.

	2008	2009	2010	2011
1	1.00000	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000	1.00000
4	0.97971	1.00000	1.00000	1.00000
5	0.97991	0.97825	1.00000	1.00000
6	0.92228	0.97896	0.97306	1.00000
7	1.00000	1.00000	1.00000	1.00000
8	1.00000	1.00000	1.00000	0.99759
9	1.00000	1.00000	1.00000	1.00000
10	1.00000	1.00000	1.00000	1.00000
11	1.00000	1.00000	1.00000	1.00000
12	1.00000	1.00000	1.00000	1.00000
13	0.77578	0.86573	0.85498	0.85272
14	0.95835	0.91794	0.96075	0.80889
15	0.99335	1.00000	0.98212	0.94971
16	0.93284	0.94755	0.92636	0.94802
17	0.82047	0.85069	0.86836	0.87255
18	1.00000	0.93041	0.94259	1.00000
19	0.93642	0.99755	0.97513	0.95442
20	1.00000	1.00000	1.00000	1.00000
21	0.93959	0.96039	0.95068	0.99920
22	0.82295	0.81349	0.79218	0.91641
23	0.85814	0.91978	0.90559	0.90006
24	0.95999	0.94998	0.94692	0.91913
25	1.00000	1.00000	1.00000	1.00000
26	0.97525	0.99823	1.00000	1.00000
27	0.86570	1.00000	1.00000	1.00000
28	0.88619	0.93828	0.95250	0.96886
29	0.99477	1.00000	0.99232	0.96075
30	1.00000	1.00000	1.00000	1.00000
31	0.94916	0.90118	0.93660	1.00000
32	0.87187	0.87185	0.84526	0.85962
33	1.00000	1.00000	1.00000	0.96798
34	0.81919	0.92667	1.00000	0.95899
35	1.00000	1.00000	1.00000	1.00000
36	0.81942	0.86890	0.82086	0.81586
37	1.00000	1.00000	0.96594	0.88977
38	1.00000	1.00000	1.00000	1.00000
39	0.73895	0.74138	0.70687	0.67025
40	1.00000	1.00000	1.00000	1.00000
41	1.00000	1.00000	1.00000	1.00000

Table A.4: Efficiency scores for the disaggregated model for 2008 to 2011.

	2008	2009	2010	2011
1	1.00000	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000	1.00000
4	1.00000	1.00000	1.00000	1.00000
5	1.00000	1.00000	1.00000	1.00000
6	1.00000	1.00000	1.00000	1.00000
7	1.00000	1.00000	1.00000	1.00000
8	1.00000	1.00000	1.00000	0.93524
9	1.00000	1.00000	1.00000	1.00000
10	1.00000	1.00000	1.00000	1.00000
11	1.00000	1.00000	1.00000	1.00000
12	1.00000	1.00000	1.00000	0.98698
13	0.98027	0.91140	0.96107	0.90910
14	1.00000	0.97390	1.00000	0.86584
15	1.00000	1.00000	1.00000	1.00000
16	0.99878	0.97603	0.95319	0.99152
17	0.83127	0.85729	0.89211	0.87490
18	1.00000	1.00000	1.00000	1.00000
19	0.99122	1.00000	1.00000	1.00000
20	1.00000	1.00000	1.00000	1.00000
21	1.00000	1.00000	1.00000	1.00000
22	0.91181	0.97809	0.98029	1.00000
23	0.91824	0.90933	0.92478	0.94778
24	1.00000	1.00000	1.00000	1.00000
25	0.69929	0.73795	0.81816	0.70101
26	1.00000	1.00000	1.00000	1.00000
27	1.00000	0.77385	0.85987	1.00000
28	0.93616	1.00000	0.99893	1.00000
29	1.00000	1.00000	1.00000	1.00000
30	1.00000	1.00000	1.00000	1.00000
31	1.00000	1.00000	1.00000	1.00000
32	0.94396	0.96355	0.99124	1.00000
33	0.98213	0.95666	0.94380	0.90563
34	0.80822	0.92070	1.00000	0.95060
35	1.00000	1.00000	1.00000	1.00000
36	0.85465	0.88314	0.82297	0.82782
37	1.00000	1.00000	1.00000	1.00000
38	1.00000	1.00000	1.00000	1.00000
39	1.00000	0.79093	0.72731	0.74221
40	1.00000	1.00000	1.00000	1.00000
41	1.00000	1.00000	1.00000	1.00000

Table A.5: Efficiency scores for the BDRG model for 2008 to 2011.

	2008	2009	2010	2011
1	1.00000	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000	1.00000
4	0.94374	0.96827	0.95391	0.98856
5	1.00000	1.00000	0.98806	0.92573
6	0.86404	0.96844	0.96710	1.00000
7	0.94199	0.99307	0.87765	1.00000
8	1.00000	1.00000	1.00000	0.96255
9	1.00000	0.91168	0.97216	1.00000
10	1.00000	1.00000	1.00000	1.00000
11	1.00000	0.96422	0.99955	1.00000
12	1.00000	1.00000	1.00000	0.97018
13	0.77991	0.85017	0.82261	0.84787
14	0.90501	0.87535	0.87902	0.76033
15	0.91856	0.99303	0.91700	0.88986
16	0.91686	0.94888	0.90029	0.91666
17	0.81239	0.85356	0.84374	0.85584
18	0.94000	0.85599	0.89835	0.91256
19	0.94189	0.99312	0.96981	0.94340
20	1.00000	1.00000	1.00000	0.97368
21	0.92474	0.95501	0.92155	0.95874
22	0.79030	0.78618	0.75636	0.85505
23	0.83983	0.92701	0.89394	0.87170
24	0.84580	0.85137	0.87307	0.83149
25	0.68012	0.72378	0.81191	0.74261
26	1.00000	1.00000	1.00000	1.00000
27	0.82089	0.77529	0.88974	0.94587
28	0.89060	0.92068	0.93925	0.95863
29	0.87136	0.92290	0.90212	0.88782
30	1.00000	1.00000	1.00000	1.00000
31	0.85034	0.82238	0.85068	1.00000
32	0.83422	0.85761	0.79717	0.82977
33	0.99467	1.00000	1.00000	0.97086
34	0.86442	0.95613	1.00000	0.97016
35	1.00000	0.95116	0.97026	1.00000
36	0.80005	0.84716	0.78448	0.77734
37	1.00000	1.00000	0.99272	0.93402
38	0.99614	1.00000	1.00000	0.99811
39	0.75645	0.76012	0.72361	0.72536
40	1.00000	1.00000	1.00000	1.00000
41	1.00000	1.00000	1.00000	1.00000

Appendix B

Combined Results

Table B.1: Combined Results for All Hospitals for 2011.

Hospital Number	Unadjusted Efficiency Score	DRG Adjusted Efficiency Score	Change in Pure Technical Efficiency	DRG Case-Mix Adjustment Factor	No Peers
1	1.00	1.00	1.00	0.84	X
2	1.00	1.00	1.00	0.69	X
3	1.00	1.00	1.00	0.59	X
4	1.00	0.95	1.01	0.59	
5	0.96	0.90	0.96	0.67	
6	1.00	1.00	1.15	0.59	X
7	0.93	0.99	1.05	0.79	
8	0.86	0.93	0.93	0.76	
9	1.00	1.00	1.00	0.69	X
10	1.00	1.00	1.00	0.90	X
11	0.81	1.00	1.00	1.00	X
12	0.79	0.95	0.96	1.09	
13	0.76	0.81	1.08	0.94	
14	0.76	0.76	0.83	0.88	
15	0.88	0.89	0.94	0.95	
16	0.89	0.92	1.01	0.98	
17	0.83	0.85	1.07	0.98	
18	0.78	0.93	0.97	1.11	
19	0.95	0.88	1.00	0.89	
20	1.00	1.00	1.00	0.95	
21	1.00	0.91	1.02	0.89	
22	0.86	0.85	1.07	0.93	
23	0.82	0.88	1.10	1.03	
24	0.92	0.80	0.95	0.84	
25	0.63	0.77	1.13	1.43	
26	1.00	1.00	1.03	1.02	X
27	0.73	0.94	1.13	1.25	
28	0.97	0.92	1.07	0.95	
29	0.96	0.88	0.98	0.91	
30	1.00	1.00	1.00	0.99	X
31	1.00	1.00	1.09	1.03	X
32	0.84	0.82	0.98	0.89	
33	0.82	0.92	0.95	1.16	
34	0.89	0.96	1.17	1.03	
35	1.00	0.98	0.98	0.93	
36	0.75	0.80	0.98	1.06	
37	0.85	0.89	0.89	1.05	
38	1.00	0.98	1.00	0.94	
39	0.66	0.67	0.92	1.02	
40	1.00	1.00	1.00	1.21	X
41	1.00	1.00	1.00	1.12	X